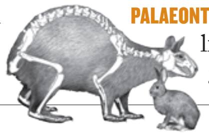
THIS WEEK

EDITORIALS



Scotland be brave

Those who want to build a better future for Scotland should resist cuts to an innovative scheme that helps its universities to compete with larger rivals elsewhere.

Not everyone in academia can have the resources of a Harvard or an Imperial College. Most university departments are modest, with constrained access to equipment and specialist expertise. A major part of academic life is the constant battle to forge the partnerships needed to deliver research and teaching of international quality.

A novel and illuminating response to this challenge began to take shape in Scotland seven years ago, with the creation of SUPA — the Scottish Universities Physics Alliance — which sought to build much closer links between the research and postgraduate teaching of physics departments at six universities. The 'pooling' idea has now been taken up by nine other disciplines — and has been widely acclaimed at home and abroad (see *Nature* **447,** 1031; 2007). Yet last month, the Scottish Funding Council (SFC), which also provides block funding to Scotland's universities, announced that its support for the pools will be cut next year by 15%.

The impact of the cuts on the pools will be especially severe because the savings will have to be found in the small portion of their budget that is not allocated to permanent staff costs — money that supports, for example, the movement of students between universities. They also illustrate a problem that is likely to resonate around the world: as government funding cuts bite, agencies, under pressure from scientists and institutions, will be tempted to pull back from funding the most innovative and arguably relevant work, and concentrate on what they view as core activities. At grant agencies such as the US National Institutes of Health, this might mean cutting interdisciplinary programmes to protect single-investigator grants. In the case of the SFC, it means maintaining the block grants that support university research, but cutting funding for the pools.

A wise agency would do the opposite. Support for Scotland's research pools should continue, even if that means finding the money by trimming the basic block grant. Cutting that whole cake would be difficult politically, because it would mean the Scottish government could not claim that it has "fully protected research funding". But it is worse to proclaim that the cake has been protected when the cream has been siphoned off. Funding the pools doesn't directly support the best research — that is backed, on the basis of the Research Assessment Exercise, by the block grants — but they are the best strategic approach for strengthening the research base.

The pools have already enhanced Scotland's strong performance in UK-wide competition for research-council grants (11% of the money for 8% of the population) and will surely help to bolster its exceptional performance in international comparisons (most citations in the world per unit of gross domestic product, and second only to Switzerland in citations per paper, according to the *International Comparative Performance of Scotland's Research Base November 2009* — a report published for the Scottish government).

The pools were hard work to set up. They had to overcome customary and powerful rivalries between neighbouring university departments. But with the support of the universities and of the SFC, the longer-established ones have developed strength and resilience. Just two weeks ago, Scottish education minister Michael Russell — who was regaled with praise for SUPA on a recent visit to CERN, the particle-physics laboratory in Switzerland — warmly complimented the pools in the Scottish Parliament, calling for them to be "taken to the next level".

"Scotland's excellent university research is one of the nation's main assets."

According to the plan subsequently released by the SFC, that 'next level' will be down. Pool leaders are fuming over the incoherence of the cuts, and university heads have called on the SFC to reallocate the relatively small amount of money involved — £3 million (US$4.8 million) — from an 'invest-to-save' component of its budget.

Some of the pools, meanwhile, will try to take their case directly to the people, who will elect a new Scottish government on 5 May. This raised profile will be welcome, as Scotland's excellent university research is one of the nation's main assets — but one rarely noticed amid the noisy debate over student fees.

Those running for the election should be asked if they would restore full funding to the research pools. They can't hide behind the SFC, the quasi-autonomous organization that makes the call, because it takes its marching orders on strategic questions from the government. Political leaders should back the research pools, which have shown themselves as a promising way to let a small country carry weight in a ferociously competitive research world. ■

Soya scrutiny

A partnership to encourage sustainable farming in Brazil may not be as green as it seems.

Large swathes of the Brazilian Amazon have come to resemble the midwestern United States in recent years, having been planted with soya as far as the eye can see. This development has unnerved conservation organizations, which fear that huge expanses of pristine rainforest are being felled to make way for the lucrative crop. A widespread hope is that large agribusiness, aware of both the need to protect this fragile environment and the importance of good public relations, can be induced to farm more sustainably. But a recent assessment of one major partnership between conservationists and the soya industry suggests the need for caution.

© 2011 Macmillan Publishers Limited. All rights reserved

In 2006, it emerged that Cargill, the US agricultural giant based in Minneapolis, Minnesota, had expanded a port in Santarém, Pará state, to handle one million tonnes of locally produced soya — without having conducted a required environmental impact assessment. (The assessment was completed in 2010.) Environmentalists cried foul, fearing the impact of expanded production on the rainforest. With the world's attention on it, Cargill formed partnerships with several green groups, including The Nature Conservancy, headquartered in Arlington, Virginia. The aim was to ensure that Cargill was purchasing sustainably grown soya from local farmers and respecting the rights and opinions of other locals who oppose the expansion of soya production.

However, Cargill's apparently environmentally friendly operations in the Amazon may not be as green as they seem. Research by Brenda Baletti, a PhD candidate at the University of North Carolina at Chapel Hill, questions claims by green groups and Cargill that its soya operations avoid deforestation. In addition, the work raises concerns over how well the operations respect local opinion. The research will be presented this week at the International Conference on Global Land Grabbing at the University of Sussex in Brighton, UK.

Baletti contends that a satellite-imaging system capable of detecting deforestation on individual farms in Santarém was not available until two years ago. Before the monitoring system was up and running, there was simply no way to judge whether soya production by particular farmers in the area was sustainable, argues Baletti. This, she says, raises doubts about claims by green groups that their initiatives to protect the rainforest have succeeded.

Baletti also questions whether the Cargill–Nature Conservancy partnership took full account of the views and concerns of small farmers and others who were opposed to industrial-scale soya production in Santarém. Interviews conducted by Baletti indicate that many local people were unhappy with the developments and felt that their perspectives were ignored. The Nature Conservancy's partnership with Cargill may not have helped much: Baletti's research highlights, for example, that the organization reports those soya farmers who breach national forest-protection laws to Cargill, but not to the government.

The Nature Conservancy told *Nature* that it has been able to monitor broader patterns of deforestation for the past six years. It regards the joint programme with Cargill as successful because it has found clear reductions in forest loss over the past three years. The organization added that working with farmers would be difficult if it reported violations of forest-protection laws to the government, and said that its approach works: more farmers are complying with the laws and sparing the forest. Cargill, for its part, says that the initiative has discouraged the planting of soya in deforested areas and has helped to bring down the rate of deforestation in Santarém. The company also says that it regularly engages with locals through consultations and public meetings, and has taken on board some of their concerns about soya farming.

"Well-intentioned schemes to protect the environment, respect local interests and boost the economy may fall short."

Yet the questions about Cargill's soya-production interests in the Brazilian Amazon echo wider concerns about whether market-driven approaches to conservation and sustainable development are always workable. Ensuring that all players, big and small, have a seat at the negotiating table is a worthwhile goal, and involving local communities is a pillar of the United Nations' programme to reduce greenhouse gases from deforestation. But it seems naive to assume that all voices and interests carry equal weight.

If not carefully designed, well-intentioned schemes that aim to protect the environment, respect local interests and boost the economy — all at once — may fall short. It should be borne in mind that if something sounds too good to be true, it often is. ■

Eye on the tiger

The latest Indian tiger census demonstrates welcome methodological rigour.

Counting wildlife is no easy matter, especially with a wide-ranging and stealthy forest dweller like the tiger, which is in the spotlight this week. On 28 March, India announced that its tiger population, in decline for decades, has increased by 225 animals over the past four years, to a total of 1,706. The scale of the increase is open to dispute, but the announcement signals a welcome development: the beginnings of a more rigorous approach to counting the big cats.

The tiger's worldwide decline is a tragedy, and nowhere is this more true than in India, which is home to more than half of the global population, and which considers the animal a national symbol and a source of pride. Independent scientists and the government have long been at odds over how to count and protect animals in the country's 39 tiger reserves.

But in 2005, the government was forced to come to the bargaining table after the embarrassing revelation that at the Sariska reserve in Rajasthan — once a stronghold for the cats — not a single tiger was left, even though the government had claimed that between 16 and 18 tigers lived there in 2004. A review of tiger management practices, headed by fierce environmental-justice campaigner Sunita Narain of the Centre for Science and Environment in New Delhi, recommended that the government revamp its tiger census, conducted once every four years.

Since then, officials and wildlife scientists have worked in a spirit of (perhaps grudging) cooperation. The government has abandoned its previous census strategy, which relied on counting tiger 'pugmarks', or tracks, and is incorporating more-modern methods, such as camera trapping and DNA testing. It has also become more serious about defending tiger reserves from encroaching development, and involving people who live near the reserves in policing them for poachers. Although long overdue, these are huge steps in the right direction.

But are they enough to have spurred such a dramatic increase in tiger numbers — especially after many years of relentless decline owing to poaching and habitat loss? It seems unlikely. Nor would it be wise to compare older tiger counts with a census taken using new methods. What matters now is to make sure that the new techniques are scientifically rigorous enough to give accurate data for future comparisons. Some information about the methodology is available, but full details have yet to be released.

Wildlife biologists such as Ullas Karanth of the Center for Wildlife Studies in Bangalore, who has brought tigers in the state of Karnataka back from the verge of local extinction, also emphasize the need to pay closer attention to tiger populations by counting them every year rather than every four years. The tiger population is in such dire straits that a single year can make a crucial difference, and limiting the count to every four years means that worrying trends might be detected too late to do anything about them.

The Indian government has signed up to the goal, and has agreed with Karanth on a methodology. Now the annual counts must begin, with total transparency about how they are done. The numbers will help to banish wishful thinking about the status of tigers. They will also help India to be sure that it is truly doing everything in its power to bring the cats back from the brink. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

© 2011 Macmillan Publishers Limited. All rights reserved

WORLD VIEW *A personal take on events*

UNIV. PORTSMOUTH

A long shadow over Fukushima

One impact of Japan's nuclear crisis is a dim but definite echo of Chernobyl, says **Jim Smith** *— decades of caesium-137.*

Three weeks after the Fukushima accident, a clearer picture is beginning to emerge of possible long-term environmental consequences. The US Department of Energy (DOE) aerial survey of radiation doses was a crucial development. A clear trace reaching out 30–40 kilometres northwest of the plant marked a zone of dose rate above 125 microsieverts per hour, a level at which immediate evacuation is often advised. Already, external doses are rapidly declining as a result of the decay of short-lived isotopes. But, as with the 1986 Chernobyl accident, it is caesium-137, with a half-life of 30.2 years, that will determine the long-term impact on the contaminated region and its residents.

The extent of caesium-137 contamination at Fukushima is not yet clear, but available data indicate very high levels in some areas. The 30 March press release from the International Atomic Energy Agency (IAEA) reports caesium-137 deposition ranging from 0.02 to 3.7 megabecquerels per square metre (MBq m^{-2}) at sites 25–58 kilometres from the Fukushima plant. The higher values are consistent with Japanese soil data from Iitate village, 40 kilometres northwest of the plant. Perhaps surprisingly, there is still no clear information on caesium-137 contamination within 20 kilometres of the plant (the distance of the evacuation zone), although the DOE map implies that this could be of the order of megabecquerels per square metre if the isotopic composition of deposits near the plant is similar to that in the area farther to the northwest.

THE SOCIAL AND PSYCHOLOGICAL RESPONSES TO RADIATION ARE OF PARAMOUNT IMPORTANCE.

The implications of these data are far-reaching. If large areas are contaminated with 0.5 MBq m^{-2} or more, evacuation could be for the long term. After Chernobyl, long-term evacuation usually occurred in areas with radioactivity above 0.55 MBq m^{-2}, although some believe that this limit could have been safely set much higher. Contamination of the food chain will depend on soil type: soils rich in clay bind radiocaesium strongly: bioavailability in organic upland and forest soils is generally significantly higher than in mineral soils. On the basis of the Fukushima data seen so far, it seems likely that in some areas, food restrictions could hold for decades (J. T. Smith *et al. Nature* **405,** 141; 2000), particularly for wild foodstuffs such as mushrooms, berries and freshwater fish.

'Liquidators' could be brought in to decontaminate towns and villages in evacuated zones and reclaim farmland, although this approach met with varying success at Chernobyl. The UK Health Protection Agency's *Recovery Handbook for Radiation Incidents* details a range of measures for residential areas, including removal of top soil and resurfacing of roads. On farms, approaches to remediation include applying potassium fertilizers to crops to compete with radiocaesium uptake, and giving 'Prussian blue' boluses to grazing animals to reduce radiocaesium absorption.

Remediation has some drawbacks: huge economic cost, for example, and potentially massive quantities of contaminated waste. Consumers may refuse products grown in contaminated areas even when they meet regulations. Chernobyl has taught us that the social and psychological responses to radiation are of great, perhaps paramount, importance.

'Headline' estimates of Chernobyl's public-health impact are dramatic: one 2006 estimate led by the International Agency for Research on Cancer foresaw 16,000 cases of thyroid cancer and 25,000 other cancers resulting from the radiation, among "several hundred million cancer cases from other causes". But risks to the individual are low. As early as 1991, an IAEA study found psychological effects to be "wholly disproportionate to the biological significance of the radiation". This study placed a high priority on providing accurate information about radiation health risks to affected populations. But 15 years later, the UN Chernobyl Forum Report still concluded that Chernobyl's impact on mental health is "the largest public-health problem caused by the accident to date". Misperceptions, and inefficient compensation, have led to widespread fatalism and feelings of victimization among locals. Resulting rises in alcohol consumption and smoking may well have done more damage than radiation exposure (see *Nature* **471,** 562–565; 2011). The failure to solve social and psychological problems relates not only to a lack of effort (at Chernobyl, vastly more has been spent on physical remediation than on public engagement), but also to the intractability of the problem.

The long-term response to Fukushima will have to be pragmatic. The Japanese authorities may have to rewrite the rule-book, as they have begun to do in allowing doses of 250 mSv for radiation workers. After an accident, it may be appropriate to set exposure limits for members of the public higher than the typical 1 mSv per year maximum. A limit of 5–10 mSv per year (perhaps with voluntary resettlement at doses above 1 mSv per year) may be appropriate, bearing in mind that millions of people in areas of high natural radioactivity worldwide are exposed to more than 10 mSv per year, and that occupational exposures (for example, to long-haul air crews) can be around 5 mSv per year.

A turning point in my understanding of Chernobyl's impacts came while studying lakes in Belarus during the mid-1990s. In an evacuated area, lake fish contained tens of thousands of becquerels per kilogram. A couple in their early seventies lived near the lake, eating the fish and growing vegetables. They were living off contaminated land, but leading the life they had chosen to lead. This wouldn't by any means be the right choice for everybody, but I am convinced they had made the right decision for them: they were Chernobyl survivors, not victims. ■

➲ NATURE.COM
Discuss this article online at:
go.nature.com/uut3pp

Jim Smith *is co-editor and lead author of* Chernobyl: Catastrophe and Consequences *(Springer, 2005). He is currently reader in environmental physics at the University of Portsmouth, UK.*
e-mail: jim.smith@port.ac.uk

© 2011 Macmillan Publishers Limited. All rights reserved

RESEARCH HIGHLIGHTS

Selections from the scientific literature

CANCER BIOLOGY

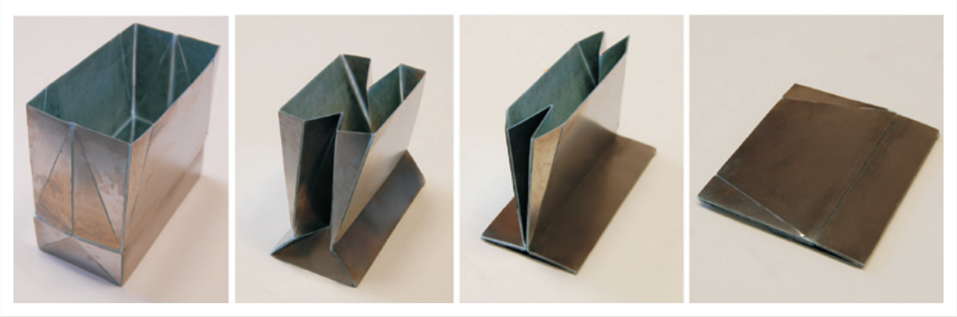
Malaria drug shrinks tumours

To grow and divide, pancreatic-cancer cells must devour their own contents — an Achilles heel that could render them susceptible to the antimalaria drug chloroquine.

'Autophagy' is the regulated degradation of cellular structures and molecules. Alec Kimmelman of Harvard Medical School in Boston, Massachusetts, and his colleagues found that pancreatic-tumour cells have high levels of autophagy. When the researchers reduced expression of the protein ATG5, which is required for autophagy, pancreatic-cancer cells showed signs of stress, including DNA damage and altered metabolism.

Furthermore, reduced expression of ATG5 or treatment with chloroquine, which inhibits autophagy, shrank tumours and lengthened survival time in mice that had received transplanted pancreatic-cancer cells.

Genes Dev. **doi:10.1101/gad.2016111 (2011)**



R. SOC.

ORIGAMI

How to fold a rigid bag

The ability to fold containers flat is invaluable to the packing industry. Packing companies mainly use either flat-packed, rigid boxes that are open at both ends and have to be fastened shut at the bottom before use, or flexible bags that can be folded flat but are not as strong.

Now Zhong You and Weina Wu at the University of Oxford, UK, have mathematically devised a hybrid solution, featuring advantages of both systems. Their complex folding pattern of 28 creases forms the design for a rigid bag with a closed bottom. They demonstrated their solution by making a bag out of paper bonded to steel sheets and folding it flat (**pictured**). This is the first time a solution has been found to fold flat a rigid bag that is taller than half its depth.

Proc. R. Soc. A **doi:10.1098/rspa.2011.0120 (2011)**

NANOTECHNOLOGY

Painting and shaping pillars

Patterning a surface with closely spaced nanometre-scale pillars can create structures that bend light, help to catalyse reactions and repel or attract fluids. But, typically, each variant pattern needs a new master mould, limiting scientists' ability to quickly test many possible patterns.

P. KIM & J. C. WEAVER, HARVARD UNIV.

Joanna Aizenberg, Philseok Kim and their colleagues at Harvard University in Cambridge, Massachusetts, now show how to quickly transform a single master array into more complicated patterns. They deposit gold or platinum onto an array of nanopillars, and use the metal-covered posts as electrodes on which to deposit conducting polymers. By varying the deposition conditions, the researchers make tapered posts (**pictured**), overhangs and other three-dimensional structures with customizable diameters and spacings.

Nano Lett. **doi:10.1021/nl200426g (2011)**

PALAEOANTHROPOLOGY

Remains in ancient cave get younger

A South African cave central to our knowledge of ancient human ancestors is not as old as previous estimates suggest. Sterkfontein Cave, near Johannesburg, has been an important source of fossil ancestors from more than 2 million years ago. But its complex geology has thwarted efforts to accurately date the fossils and artefacts buried there.

Andy Herries at the University of New South Wales in Sydney, Australia, and John Shaw at the University of Liverpool, UK, recorded the geomagnetic orientation of different layers of calcite deposits. Knowing when during history Earth's magnetic field reversed provided a basic chronology. From this, the researchers established that many of the deposits, and hence the cave's human remains, are younger than previously thought.

For instance, an *Australopithecus* fossil known as Little Foot is less than 2.6 million years old — much younger than the well known 3.2-million-year-old *Australopithecus afarensis* fossil from Ethiopia named Lucy.

J. Hum. Evol. 60, **523–539 (2011)**

ASTRONOMY

Comets gave rings their ripples

Ripples in the rings of Saturn and Jupiter are the consequence of collisions with cometary fragments, according to two papers.

Matthew Hedman at Cornell University in Ithaca, New York, and his colleagues spotted the ripple in one of Saturn's rings while analysing images taken by the Cassini

© 2011 Macmillan Publishers Limited. All rights reserved

spacecraft in 2009. Because a ripple's wavelength is related to how long ago it was created, the researchers calculated that this ripple began in 1983. They say that debris from a colliding comet crashed into particles in the ring, causing it to tilt and wobble.

Another group that included Hedman and was led by Mark Showalter at the SETI Institute in Mountain View, California, found two sets of ripples in images of Jupiter's rings taken by the Galileo and New Horizons spacecraft. The team reports that Jupiter's rings were hit by debris from two comets, the first of which struck in 1990 and the second in 1994. The authors associated this latter hit with a known Jupiter strike by the Shoemaker–Levy 9 comet in July 1994.

Monitoring the rings' wobble may help astronomers to map the internal structure of the planets' cores.

Science **doi:10.1126/science.1202238; doi:10.1126/science.1202241 (2011)**

For a longer story on this research, see go.nature.com/a5qsys

CHEMISTRY

Fine-tuning polymerization

Synthetic polymers are all around us. Researchers are working on ways to more finely control the chemical process that produces these compounds, to generate polymers with, for example, specific molecular masses and architectures.

Krzysztof Matyjaszewski at Carnegie Mellon University in Pittsburgh, Pennsylvania, and his collaborators show that applying a suitable electrochemical potential to a copper catalyst commonly used in polymerization reactions switches the catalyst between a dormant (oxidized) and an active (reduced) state. Because the polymerization reaction rate depends on the ratio of the concentrations of the two copper species, this toggling provides fine control over the yield and molecular mass of the product. Moreover, polymerization can be stopped and restarted at will by adjusting the applied potential.

The approach may be extendable to various types of polymer, and might be used to generate complex architectures, the authors say.

Science 332, **81–84 (2011)**

EVOLUTIONARY BIOLOGY

Bad for your teeth?

Faecal analysis has added grist to the debate over whether dietary silica contributed to the evolution of herbivores' high-crowned teeth. The compound, which is found in dust and in certain plants, such as grasses, is harder than tooth enamel and can wear teeth down.

Jürgen Hummel at the University of Bonn in Germany and his colleagues measured the levels of silica, which is not absorbed or degraded during digestion, in the faeces of 15 species of African herbivore, including giraffes, antelopes and zebras. They used this as a proxy for the amount of silica ingested, and found that herbivores with higher-crowned teeth tend to have higher levels of silica in their diet. This supports the argument that herbivores evolved these teeth to offset the dental erosion caused by their diet.

Proc. R. Soc. B **doi:10.1098/rspb.2010.1939 (2010)**

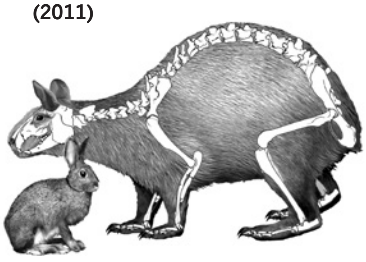
PALAEONTOLOGY

The giant rabbits of Minorca

Fossil bones from a species of giant rabbit, *Nuralagus rex*, have been discovered on the Mediterranean island of Minorca, off the coast of Spain.

Meike Köhler at the Catalan Institute for Research and Advanced Studies in Barcelona, Spain, and her colleagues dated the fossils to the late Neogene period, about 5 million years ago. The outsized rabbit (reconstruction **pictured**) would have weighed about 12 kilograms — 10 times the weight of its closest living continental relative, *Oryctolagus cuniculus* (**pictured**). The hefty animal also had a relatively small skull and sensory organs, and lacked the backbone anatomy required for hopping — traits, the authors say, that may have evolved because the rabbits had no predators on the island.

J. Vertebr. Paleontol. 31, **231–240 (2011)**



SOC. VERTEBR. PALEONTOL.

COMMUNITY CHOICE

The most viewed papers in science

CELL BIOLOGY

Starved cells opt for quiescence

HIGHLY READ on jcb.rupress.org in March

Many cells move in and out of a 'quiescent' state, in which they stop dividing. Researchers in France show that the deprivation and addition of glucose triggers yeast cells to enter and exit from this state, respectively, regardless of which stage of the cell cycle they are in. This contradicts the common belief that cells can become quiescent only when they are in a specific phase of the cell cycle.

Isabelle Sagot and her colleagues at the National Centre for Scientific Research in Bordeaux found that *Saccharomyces cerevisiae* cells in all stages of the cell cycle display the cellular hallmarks of quiescence entry. The authors also found that glucose added to the cells had to be metabolized to a certain point for the cells to exit from quiescence. This suggests that quiescence signals are more closely linked to the cell's metabolic status than to the cell cycle.

J. Cell Biol. 192, **949–957 (2011)**

FLUID DYNAMICS

Getting through a drowning machine

People adrift in fast-moving rivers can get caught and even drown in hydraulic jumps — turbulence that forms when a shallow, fast-moving waterway spills into a slower and deeper section. By creating a small hydraulic jump in the lab, researchers have determined the factors that affect how long different objects remain caught in these 'drowning machines.'

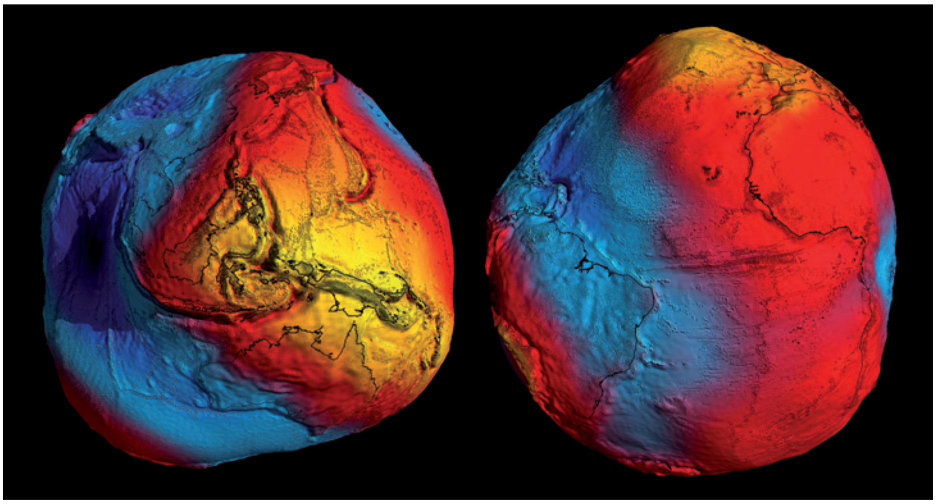
Pinaki Chakraborty and his colleagues at the University of Illinois in Urbana observed the time it took objects of different shapes and buoyancies, such as a ball and a bottle, to emerge from the jump. Using these data, they calculated that light objects remain in the rolling eddy for several minutes, whereas denser items — including humans — should be expelled after only seconds. Because object buoyancy affected residence time, the authors suggest that public-safety agencies test the best strategy for people to use if they fall into a drowning machine: remaining still, and thus more buoyant, or swimming with the current to emerge from the jump sooner.

Proc. Natl Acad. Sci. USA **doi:10.1073/pnas.1015183108 (2011)**

➲ NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

© 2011 Macmillan Publishers Limited. All rights reserved

SEVEN DAYS *The news in brief*



ESA/HPF/DLR

The pull of the planet

The most detailed map of Earth's gravity ever made was unveiled last week in Munich, Germany, when researchers presented eight months' worth of data from the European Space Agency's Gravity Field and Steady-State Ocean Circulation Explorer (GOCE), a satellite launched in 2009. GOCE maps subtle variations in Earth's gravitational field that arise from the planet's uneven distribution of mass. The result is a 'geoid' (pictured — variations exaggerated 10,000 times), showing the world if it were covered by an ocean whose height was influenced only by gravity. This reference allows geoscientists to precisely measure the heights of shifting oceans and continents. GOCE will continue mapping until the end of 2012.

EVENTS

Nuclear struggles

Radioactive water flooding turbine buildings at Japan's stricken Fukushima nuclear power plant continued to be the main concern for the plant's operator, the Tokyo Electric Power Company (TEPCO), last week. On 4 April, TEPCO said that it was deliberately pumping contaminated water into the sea, to free storage space for water with higher levels of radioactivity. External experts are growing increasingly mistrustful of information provided by TEPCO; Japan's Nuclear and Industrial Safety Agency criticized the company for providing inaccurate radiation data. See pages 13–14 for more on the state of the reactor and details of the fallout.

POLICY

Rare diseases

A global collaboration that aims to find new therapies for rare diseases was launched this week by the US National Institutes of Health and the European Commission. The International Rare Disease Research Consortium wants to develop a diagnostic tool for every known rare disease by 2020, and find therapies for 200 of them. The project will involve research agencies from around the world. See page 17 for more details.

French stimulus

France's medical-research landscape has been reshaped by the government's 30 March announcement of six new centres of excellence, designed to promote research bridging the gap from bench to clinic. The centres bring together scientists from universities and hospitals to focus on specific fields such as cardiovascular disease, infectious diseases and neuroscience. They will share an €850-million (US$1.2-billion) fund — much of it in endowments intended to provide support for the next decade. The funding is part of the first wave of a €35-billion economic stimulus package announced in December 2009. See go.nature.com/vydpc9 for more.

Gene patents

An appeals court in Washington DC heard pivotal arguments on 4 April in a landmark case with far-reaching implications for gene patenting. Last year, a federal judge in New York ruled that seven patents on the breast-cancer genes *BRCA1* and *BRCA2* held by Myriad Genetics and the University of Utah Research Foundation, both based in Salt Lake City, were "improperly granted" because they related to a product of nature (and therefore were not patentable). The appeals-court ruling is expected in June, but many think an appeal to the supreme court will follow. See go.nature.com/bxcmmj for details.

PEOPLE

Templeton prize

Astrophysicist Martin Rees has won this year's £1-million (US$1.6-million) Templeton Prize. The award is granted for those who have made an exceptional contribution to "affirming life's spiritual dimension". Rees, former head of the Royal Society in London and an emeritus professor at the University of Cambridge, UK, has worked on areas such as galaxy formation, black holes and γ-ray bursts. Some researchers and atheists are unhappy with the increasing tendency of the Templeton Foundation — based near Philadelphia, Pennsylvania — to select scientists for its awards and grants (see *Nature* **470,** 323–325; 2011). See go.nature.com/ub2ett for more on the prize.

Relics boss returns

Archaeologist Zahi Hawass has been reinstated as Egypt's antiquities minister, less than a month after quitting the position. Hawass, a flamboyant figure who is often in the media, had stepped down on 3 March, claiming that security services couldn't adequately

© 2011 Macmillan Publishers Limited. All rights reserved

protect the archaeological sites and museums under his jurisdiction. Essam Sharaf, prime minister of Egypt's interim government, reappointed him on 30 March. See go.nature.com/1emirz for more.

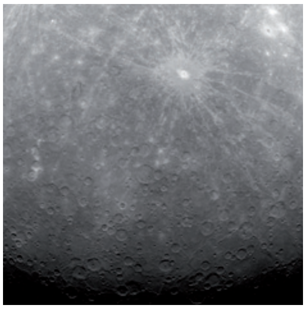
RESEARCH

Arctic ozone

Stratospheric ozone loss in the Arctic has this year reached a level never before recorded in the Northern Hemisphere. Observations made since the beginning of the Arctic winter show that 40% of ozone molecules have been destroyed over the Arctic. The highest ozone loss previously measured was 30%, in 2005. Scientists with the World Meteorological Organization and the Alfred Wegener Institute for Polar and Marine Research released the figures on 5 April at the general assembly of the European Geosciences Union in Vienna. See go.nature.com/dxmamu for more.

Mercury mapping

NASA's MESSENGER spacecraft has sent back the first images ever taken from orbit around Mercury. The very first image (**pictured**) shows a bright, rayed crater named Debussy near the planet's south pole. Mission scientists think that the shadowed craters south of Debussy might harbour permanent water ice. MESSENGER — which entered into orbit around Mercury on 18 March — has now started a planet-wide mapping survey. See go.nature.com/rdqsni for more.



NASA/JOHNS HOPKINS UNIV. APPL. PHYS. LAB./CARNEGIE INST. WASHINGTON

BUSINESS

Drug discovery

Yale University's school of medicine is to collaborate with biotechnology firm Gilead Sciences to discover cancer therapies. Gilead, based in Foster City, California, will provide US$40 million for a four-year research effort, and may extend that to $100 million over ten years. The company is best known for selling HIV drugs, but has made three cancer-related acquisitions in the past year. The partnership with Yale, of New Haven, Connecticut, was announced on 30 March, and continues a trend of drug-discovery collaborations between industry and academia. The University of California, San Francisco, has announced partnerships with both Sanofi-aventis of Paris and Pfizer of New York in the past six months, for instance.

Pfizer divesting

Pharmaceutical giant Pfizer may be preparing to shed some of its business units after its US$68-billion 2009 mega-merger with Wyeth. On 4 April, the company, headquartered in New York, said that it had agreed to sell its capsule-manufacturing division, Capsugel, to Kohlberg Kravis Roberts, a private-equity firm also in New York, for $2.4 billion. Pfizer has already said that it is reviewing its portfolio; analysts speculated that this could herald a series of sales of units not directly related to pharmaceuticals, such as nutrition or consumer-health divisions.

Hostile bid

On 29 March, Valeant Pharmaceuticals issued a US$5.7-billion hostile takeover bid to acquire Cephalon, the maker of the wakefulness drug modafinil. Cephalon, based in Frazer, Pennsylvania, is best known for its pain and sleep drugs, but has recently diversified its business, acquiring companies that fortified its portfolio in cancer and generic pharmaceuticals. Valeant Pharmaceuticals is based in Mississauga, Canada.

Chemical deals

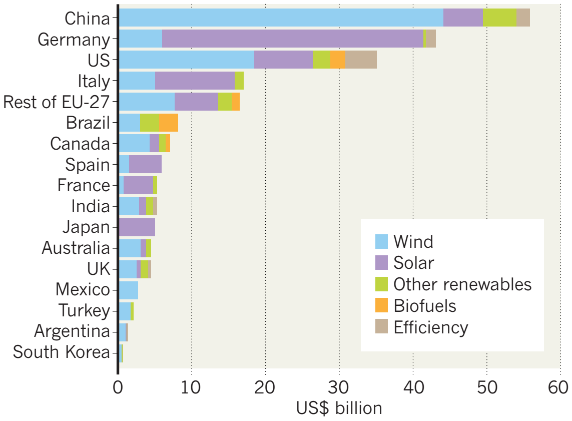
In the latest in a flurry of chemical deals, the Belgian chemicals group Solvay has agreed to buy its smaller French rival Rhodia for €3.4 billion (US$4.8 billion), the two firms said on 4 April. Solvay sold its pharmaceutical business to US drug firm Abbott Laboratories last year for €4.5 billion, and had been looking to reinvest. Well over $20 billion has been bid in acquisition offers in the chemicals sector so far this year.

TREND WATCH

China last year produced almost half of the world's wind turbines and solar modules, and attracted a record US$54.4-billion investment in clean energy (excluding nuclear power), says a report released last week by the Pew Environment Group in Washington DC. The country-by-country breakdown (see chart) excludes government stimulus funding and global research investments of around US$35 billion. Germany invested most relative to its economy, at 1.4% of gross domestic product.

CHINA LEADS CLEAN ENERGY PACK

China alone attracted a quarter of non-governmental investment in clean energy in 2010.



SOURCE: PEW ENVIRONMENT GROUP/BNEF

COMING UP

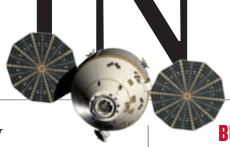
9–13 APRIL
Around 13,000 scientists are expected at the Experimental Biology 2011 conference in Washington DC. The meeting covers research from fields such as anatomy, biochemistry and pharmacology.
www.experimentalbiology.org

12 APRIL
More than 300 events are planned across the world to celebrate the 50th anniversary of the first human spaceflight by Soviet cosmonaut Yuri Gagarin. It is also the 30th anniversary of the first NASA shuttle flight.
http://yurisnight.net

11 APRIL
A landmark assessment of nitrogen flows across Europe, and how the element affects soil, water and biodiversity, is presented at a workshop on nitrogen and global change running until 14 April in Edinburgh, UK.
www.nitrogen2011.org

➲ NATURE.COM
For daily news updates see:
www.nature.com/news

© 2011 Macmillan Publishers Limited. All rights reserved



REUTERS/KYODO

Children are particularly at risk of thyroid cancer from radioactivity in contaminated food and drink.

NUCLEAR ACCIDENT

Fukushima health risks scrutinized

But scientists are struggling to pick through radiation data.

BY DECLAN BUTLER

Even as the damaged reactors at the Fukushima Daiichi nuclear power station continue to leak radiation, researchers have begun laying the groundwork for studies that will look for any long-term effects on public health.

Academic scientists face major obstacles as they try to collate baseline data on radiation doses in the face of the enormous disruption caused by the earthquake and tsunami that hit the country last month. But the experience of the 1986 Chernobyl nuclear accident shows that such baseline data are vital. Without them, drawing firm conclusions about any adverse health effects will be much more difficult.

Researchers emphasize, however, that environmental levels of radiation outside the 20-kilometre evacuation zone around the power plant are currently far below levels that warrant concerns about human health. The greatest threat to human health from the disaster is consuming contaminated food and drink, they say.

Assessing the impact of any exposure to radiation in the environment may require cohort studies to look for a raised incidence of cancers years from now among people living in regions with raised levels of contamination. Just how far-reaching those studies need to be, or whether they are needed at all, will depend on the extent of the ongoing contamination from the damaged reactors. Although the prevailing winds are blowing the bulk of radioisotopes from the plant out over the Pacific Ocean, periodic changes in weather patterns are dumping fallout inland, increasing the doses that residents receive. The Japanese authorities acknowledged last week that it may be months before the reactors are brought under control. For now, “it is difficult to predict what the health effects might be”, says Dillwyn Williams, a cancer researcher at the Strangeways Research Laboratory in Cambridge, UK.

Plant workers are being exposed to much higher levels of radiation than the general population, and will be monitored for long-term health effects. The Tokyo-based Radiation Effects Association already has an ongoing study of the health of Japanese nuclear-power workers, and new dosimetric data for Fukushima workers will be merged into that study.

But the Radiation Effects Research Foundation (RERF), based in Hiroshima and Nagasaki, which is responsible for radiation epidemiology studies on survivors of atomic-bomb explosions, is already initiating discussions on broader Fukushima studies. In a joint statement to *Nature* (see go.nature.com/cckfoe), the RERF’s vice-chairman and chief of research Roy Shore, and Kotaro Ozasa, its head of epidemiology, say that it is vital to gather baseline data — such as the exact locations of people exposed to fallout — as soon as possible. Several agencies, including Japan’s science ministry, local authorities and the Tokyo Electric Power Company, the plant’s operator, are already publishing measurements. But compiling and evaluating the information will be a challenge, say Shore and Ozasa, as these data are currently “scattered and uncoordinated”.

“The problem is that it is very difficult to get a real picture of the exposure of the population,” says Elisabeth Cardis, a radiation epidemiologist at the Centre for Research in Environmental Epidemiology in Barcelona, Spain. A critical review of all the available data is desperately needed, she says (see go.nature.com/ejlpny).

Questionnaires should also be sent to people in higher-risk areas to identify details such as the time spent outdoors on various dates, and what food and water were consumed, say Shore and Ozasa. “Obviously, it is important to obtain those data sooner rather than later, but at this point, coping with the huge effects of the earthquake and tsunami has to take precedence,” their statement says.

Japan’s prompt evacuation of the 20-kilometre zone around Fukushima, and bans on suspect produce, should have helped to curb exposure to isotopes of concern. Iodine-131, which ▶

© 2011 Macmillan Publishers Limited. All rights reserved

▶ has a half-life of just 8 days but accumulates quickly in the thyroid gland, is still the major component of the emissions from the nuclear plant and remains the greatest acute radiation health threat to the public, says Richard Wakeford, an epidemiologist at the Dalton Nuclear Institute, University of Manchester, UK. Some of the radioactivity levels detected in food since the accident have been "pretty hefty", he adds.

One of the largest health impacts from Chernobyl has been the 6,000 or more cases of thyroid cancer, mostly affecting people who were children at the time of the accident. In most of these cases, people received high radiation doses through drinking milk from cows that had grazed on iodine-contaminated pasture. Children are particularly at risk because their thyroids are still developing and are more prone to radiation damage than adults' mature thyroids.

The Japanese authorities are distributing potassium iodide tablets in affected areas, and Williams says that this is a crucial precaution. The tablets swamp the thyroid with non-radioactive iodine, preventing uptake of the radioactive form. Japanese children may also have a cultural advantage that lowers their risk from radioiodine. Whereas the children of Chernobyl tended to be iodine-deficient, the Japanese diet, rich in fish and seaweed, is "one of the most iodine-rich diets in the world", says Williams. Milk is also far less important in the Japanese diet than it was for the rural populations around Chernobyl.

> **"The problem is that it is very difficult to get a real picture of the exposure of the population."**

Radioiodine doses in the thyroids of children in the most contaminated areas are already being monitored by the Japanese authorities. *Nature* has learned that the first results of that survey show minimal thyroid doses in 946 children living in areas northwest of the plant, a region where some of the highest fallout over land has been reported. Measurements during 28–30 March found maximum doses of 0.07 microsieverts per hour. This would suggest that the children had received total doses of less than 100 microsieverts, many thousands of times lower than was received by people living in contaminated areas around Chernobyl. The results "seem reassuring that not much iodine-131 has got into children", says Wakeford, adding that if the food bans are being effective, "Japan will have got a grip on what is the major concern in this sort of situation".

Vadim Chumak, a health physicist at the Research Center for Radiation Medicine at the Academy of Medical Sciences of Ukraine in Kiev, who has coordinated Chernobyl health studies, says that Japanese radiation researchers should heed a key lesson from that disaster. Dose data are fleeting, he warns, and if they are not collected now, any eventual research would be much more prone to uncertainty. Dosimetric monitoring after Chernobyl was sub-standard, he says, "so in our research we had to invest enormous time and effort in the retrospective estimation of doses". ■

NUCLEAR ACCIDENT

Japan's long road ahead

Isotopes hint at more than a decade of clean-up at Fukushima.

BY GEOFF BRUMFIEL

It came as no surprise when the Tokyo Electric Power Company (TEPCO) admitted last week that it will scrap its stricken Fukushima Daiichi reactors. After explosions, copious radioisotope leaks and a liberal dousing with sea water, the reactors are a write-off. But what will workers encounter when they finally start decommissioning the shattered plant?

On 11 March, a tsunami knocked out backup generators, preventing cooling water from circulating around the hot cores of reactors 1–3. The fuel rods inside began to warp, split and at least partially melt. Steam reacted with the rods' outer sheath of zirconium, creating hydrogen gas that caused a sequence of explosions (see *Nature* **471,** 417–418; 2011).

But data from Japanese regulators and TEPCO suggest to some researchers that conditions inside the core could be far worse than a partial meltdown. Some believe that molten fuel may have flowed into the outer concrete containment vessel, whereas others suggest that nuclear chain reactions are still happening inside the fuel.

The damaged Fukushima Daiichi power plant.

AP PHOTO/AIR PHOTO SERVICE

➲ NATURE.COM
For a longer version of this story see:
go.nature.com/octevi

The most worrisome evidence comes from water found in a building next to reactor 1. On 26 March, Japan's Nuclear and Industrial Safety Agency reported the presence of chlorine-38, a radioisotope with a half-life of just 37 minutes that forms when natural chlorine-37 is hit by neutrons from fission. This could be evidence that fuel has clumped together into sufficiently large chunks to briefly restart nuclear reactions, says Ferenc Dalnoki-Veress, a physicist at the Monterey Institute of International Studies in California. Such bursts could put workers at extreme risk of radiation exposure during clean-up, he warns, and seriously complicate work at the site.

But Paddy Regan, a nuclear physicist at the University of Surrey in Guildford, UK, is sceptical. Other radioisotopes have a similar γ-ray spectrum and could be mimicking chlorine-38, he says, and TEPCO has already retracted erroneous measurements of other isotopes. Dalnoki-Veress agrees that the evidence is circumstantial, adding that he is frustrated by the lack of clear data coming from the plant.

Other theories also rest on tentative evidence. Richard Lahey, an emeritus professor of nuclear engineering at Rensselaer Polytechnic Institute in Troy, New York, believes that the core of reactor 2 may have melted its control rods, which are designed to stop nuclear reactions. Provisional pressure readings and high levels of radioactivity suggest to him that molten fuel has flowed through the control-rod system like lava and dripped into the containment vessel below, creating a clean-up nightmare.

The confusion recalls the weeks that followed the partial meltdown of a reactor at Three Mile Island in Pennsylvania in 1979. In the immediate aftermath of that emergency, the state of the reactor core was subject to "an ongoing debate that went on for months", recalls Jack DeVine, an independent nuclear consultant who spent six years cleaning up that accident.

Many scientists believed that the fuel rods at Three Mile Island were more-or-less intact, on the basis of computer models and simulations, says DeVine. But when a camera was finally lowered into the core in 1982, the damage was far worse than anyone had predicted. "It looked like my gravel driveway — a mess," he says. Engineers hoping to remove fuel rods in a process akin to conventional decommissioning of a nuclear core had to rethink their strategies.

It took 14 years to clear most of the fuel out of the reactor at Three Mile Island. Based on what he has seen so far, DeVine believes that decommissioning Fukushima will probably take longer. ■

© 2011 Macmillan Publishers Limited. All rights reserved

NUCLEAR ENERGY

US radiation study sparks debate

Researchers divided on how best to probe any possible link to cancer.

BY GWYNETH DICKEY ZAKAIB

VENTURE MEDIA GROUP/AURORA PHOTOS/CORBIS

Some studies have found links between childhood cancer and proximity to power stations.

Japan's ongoing nuclear emergency has intensified discussion on a simmering issue: the potential cancer risk from living near a reactor that is operating normally.

Last year, long before the crisis in Japan, the US Nuclear Regulatory Commission (NRC) asked the National Academy of Sciences (NAS) to examine this cancer question, prompted in part by long-standing public unease. The NAS is now consulting with experts about how to design a study, with the next public meetings on the effort scheduled for 18–19 April in Chicago, Illinois. Already, however, some researchers have questioned the study's feasibility and expressed doubt over whether it will produce meaningful results.

According to the NRC, less than 1% of a person's total annual background-radiation exposure comes from living near nuclear power plants. Much more comes from natural sources in the earth and air, and from some medical exams. Even so, "there are recurrent concerns among the public about increased cancer risks", says Terry Brock, the NRC's project manager for the Analysis of Cancer Risk in Populations Near Nuclear Facilities study. "We want the most current and most scientifically valid information to respond."

The last US-wide study, which found no evidence of a problem, was published by the National Cancer Institute in 1990. Now the NRC aims to update this effort by taking advantage of two decades of improvements in data and technology. For example, whereas the 1990 study considered only cancer deaths, better record-keeping means that researchers can now look for suspect patterns in cancer diagnoses. The previous study also lumped people by county, regardless of their actual distance from a nuclear plant. Global positioning systems, which can pinpoint where people live in relation to a reactor, should now help provide more meaningful results. A further step would be including estimates of radiation doses and looking for correlations with cancer incidence.

"They may make the public feel better, but they're not going to see very low-dose effects."

But Edward Maher, president of the US-based Health Physics Society, says that even if the study takes all of those factors into account, its statistical power will be too low.

"We feel that those studies don't have a lot of value," says Maher. "They may make the public feel better, but they're not going to see very low-dose effects." The money would be better spent on more laboratory research, he adds, where confounding factors such as the presence of other carcinogens can be effectively controlled.

Other experts say that the NAS should build on and improve a 2008 German study (C. Spix *et al. Eur. J. Cancer* **44,** 275–284; 2008), which found a roughly 1.5-fold increase in cancers in children younger than 5 living within 5 kilometres of nuclear power plants. The authors concluded that plant emissions were too low to explain the effect, and similar studies done later in France and Britain failed to show any cancer increase, but some researchers have challenged their interpretation of the data.

Nevertheless, Steve Wing, an epidemiologist from the University of North Carolina at Chapel Hill, says that if there is an effect, it will be easiest to see in children and fetuses. Their rapidly dividing cells make them more sensitive to radiation than adults, and they haven't been exposed to as many possible carcinogens. Wing and his colleagues wrote an article on how best to design the NAS study in the 1 April issue of *Environmental Health Perspectives* (S. Wing *et al. Environ. Health Perspect.* doi:10.1289/ehp.1002853; 2011). Among other things, they emphasize the need to obtain radiation-dose estimates for the populations under study.

In the upcoming April meetings, the NAS committee will discuss nuclear power plant emission monitoring and hear study design suggestions. After a series of additional meetings, the committee aims to complete recommendations by the end of 2011, after which they will be posted online for public comment. If the committee decides to move forward with the study, another committee will be appointed next year to carry it out.

Some experts think that there is no effect for the study to find. Antone Brooks, a radiation toxicologist at Washington State University Tri-cities in Richland, says that DNA repair mechanisms and selective suicide of damaged cells are adequate to handle DNA damage below a certain dose threshold.

"We've lived in a sea of radiation throughout evolution," says Brooks. "The body knows how to handle low doses."

Others believe that the risk never vanishes. DNA repair mechanisms don't work perfectly 100% of the time, and even small amounts of radiation confer some risk, says Bill Morgan, the director of radiation biology and biophysics at Pacific Northwest National Laboratory in Richland. "It's a tremendous debate," he says.

Some will argue that if no effect is found, there isn't a problem, says David Brenner, director of the Center for Radiological Research at Columbia University in New York. "But the fact that you can't measure a risk in an epidemiological study doesn't mean that the risk isn't there." ■

© 2011 Macmillan Publishers Limited. All rights reserved

SPACE EXPLORATION

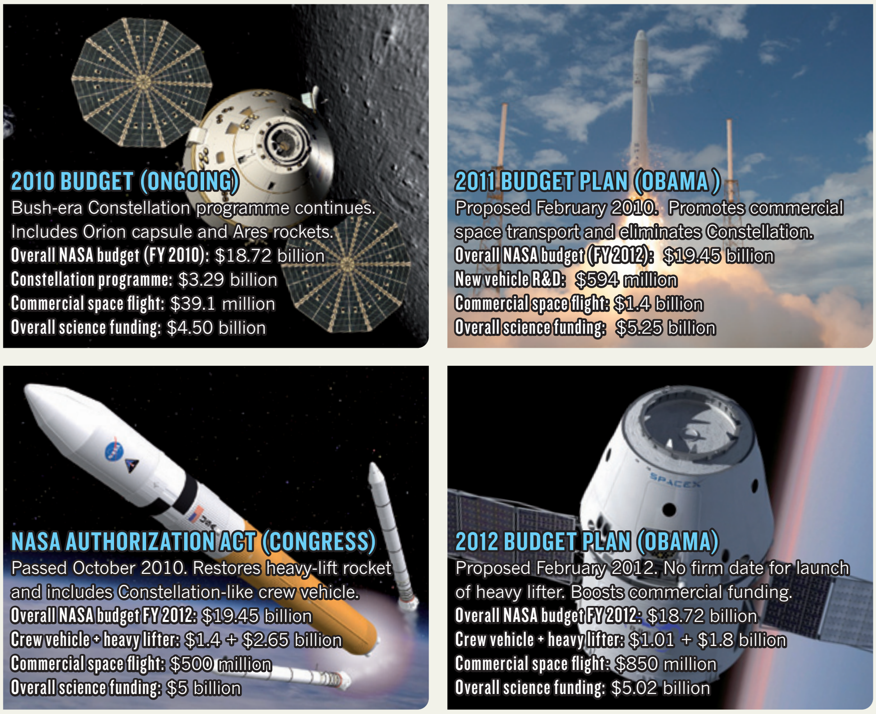
NASA human space–flight programme lost in transition

US space agency is wrestling with competing visions and uncertainty of budget deadlock.

BY ADAM MANN

SPACE WARS

NASA is currently bound to continue funding the cancelled Constellation programme. President Obama wants to shift low–Earth–orbit launches to commercial companies such as Space X (maker of the Falcon 9 rocket and Dragon capsule, pictured right, top and bottom), whereas Congress wants NASA to develop its own vehicles (such as the Orion capsule, top left, and the cancelled Ares V rocket, bottom).



2010 BUDGET (ONGOING)
Bush-era Constellation programme continues. Includes Orion capsule and Ares rockets.
Overall NASA budget (FY 2010): $18.72 billion
Constellation programme: $3.29 billion
Commercial space flight: $39.1 million
Overall science funding: $4.50 billion

2011 BUDGET PLAN (OBAMA)
Proposed February 2010. Promotes commercial space transport and eliminates Constellation.
Overall NASA budget (FY 2012): $19.45 billion
New vehicle R&D: $594 million
Commercial space flight: $1.4 billion
Overall science funding: $5.25 billion

NASA AUTHORIZATION ACT (CONGRESS)
Passed October 2010. Restores heavy-lift rocket and includes Constellation-like crew vehicle.
Overall NASA budget FY 2012: $19.45 billion
Crew vehicle + heavy lifter: $1.4 + $2.65 billion
Commercial space flight: $500 million
Overall science funding: $5 billion

2012 BUDGET PLAN (OBAMA)
Proposed February 2012. No firm date for launch of heavy lifter. Boosts commercial funding.
Overall NASA budget FY 2012: $18.72 billion
Crew vehicle + heavy lifter: $1.01 + $1.8 billion
Commercial space flight: $850 million
Overall science funding: $5.02 billion

IMAGE LEFT, C. THOMPSON/SPACE X; FAR LEFT, NASA
IMAGE LEFT, SPACE X; FAR LEFT, NASA

NASA should be revitalized "not just with dollars, but with clear aims and a larger purpose," US President Barack Obama said last April, after cancelling the previous administration's under-resourced Constellation programme of rockets and capsules for human space flight. But 12 months later, money and clarity are in short supply at the agency, which finds itself hamstrung by a budget showdown and buffeted by conflicting messages from Obama and the US Congress about the next steps in human space flight (see 'Space wars').

This week may finally bring some relief on the fiscal front — if Congress manages to pass a 2011 budget, now more than six months overdue. That would free NASA from its lingering 2010 budget requirements, which have prevented it from starting new projects or from terminating funding for Constellation. But a larger debate remains unresolved over what rockets to build to replace the venerable space shuttle, due to make its penultimate flight on 29 April.

"I've spent over 40 years closely observing the US space programme and I've never seen it as confused as it is now," says John Logsdon, former director of the Space Policy Institute at George Washington University in Washington DC. "It's simply a mess."

The goal of the Bush-era Constellation programme was to develop rockets and capsules that could both replace the shuttle and take astronauts beyond low-Earth orbit to the Moon and, ultimately, to Mars. In cancelling it, Obama called for increased spending on new rocket technologies for voyages beyond Earth's orbit, extended the end date of the International Space Station (ISS) from 2015 to 2020, and invested in private space-flight companies to ferry crew and supplies to low-Earth orbit.

The plan failed to impress Congress, particularly those members representing regions that benefit from the federal dollars NASA contracts bring. It "has spawned thousands of lost jobs" and "cast fear and doubt" throughout the space-flight industry, said Ralph Hall (Republican, Texas), chairman of the House Committee on Science, Space, and Technology, during a House subcommittee hearing on human space exploration on 30 March. Other legislators caution that Obama's proposal to buy space transportation services from private contractors is an invitation to delay and possibly disaster.

In September 2010, Congress offered up a proposal to resurrect parts of Constellation under another name. In an authorization bill — which provided direction but no money — it told NASA to produce a multi-purpose crew vehicle (MPCV) and a heavy-lift launch system with similar specifications to those of Constellation's Orion crew capsule and Ares V rocket. The new rocket, to launch by 2016, would have to be capable of taking astronauts beyond low-Earth orbit. The bill calls for NASA to maintain as many contracts from Constellation as possible to avoid the US$2.5 billion in termination fees that Obama's plan would have triggered. The plan's overall budget comes in about $1 billion lower than projected for Constellation.

➲ NATURE.COM
For more on the US budget crisis see:
go.nature.com/eg2noo

Obama signed the bill into law in October, but the debate is far from settled. In its 2012 budget request, the administration allocates $850 million towards aggressively promoting the development of commercial space transportation — 70% more than Congress authorized. And although Obama's request includes $2.81 billion for work on the MPCV and heavy-launch vehicle, it does not specify a target date for the launch of the rocket, or what vehicle might carry the MPCV in the interim.

The Obama administration and Congress also differ on what size the heavy-lift rocket should be. The authorization act says it must be capable of delivering 130 tonnes into orbit, 4.5 times more than the shuttle. Last month, NASA administrator Charles Bolden told reporters that he does not think that the 130-tonne lift capability is necessary for at least a decade, when the president's plan calls for

© 2011 Macmillan Publishers Limited. All rights reserved

manned missions beyond low-Earth orbit. Doug Cooke, NASA's associate administrator for exploration systems, expanded on this during the House subcommittee hearing, saying that NASA officials plan to develop a vehicle with an initial capability of 70–100 tonnes, which would allow the agency to launch it by 2016 or soon after.

Meanwhile, funding remains in limbo. The previous Congress failed to pass a budget for this year, and November's mid-term elections swept a Republican majority into the House of Representatives that is bent on making drastic cuts to government spending. With the two parties deadlocked over the 2011 budget, members have had to opt for a series of short-term measures that maintain 2010 programmes and funding levels. Constellation continues to be funded, delaying work on any new initiative.

Some argue that the roughly $250 million spent on Constellation in the current fiscal year has not been wasted. For example, on 21 March, Lockheed Martin Space Systems, based in Denver, Colorado, unveiled a new simulation centre where engineers will try out docking manoeuvres with the programme's Orion crew capsule. The capsule meets most requirements from the authorization bill and all indications are that it will be selected as the MPCV, says Larry Price, deputy programme manager for Orion. Continuing Constellation's contracts in this way is in NASA's best interests, he says. "As you can imagine, it would have been hugely inefficient to stop something, redistribute the labour force, and start it over again — especially if it's exactly the same," he says.

Even if NASA finally achieves the clarity Obama promised a year ago, it faces many years with no way to send people into space. The last time the agency had a similar gap — between the end of the Apollo programme in 1975 and the first shuttle launch in 1981 — it knew what was to come next. The shuttle programme had been announced three years before Apollo's conclusion.

The current situation is much worse, said James Maser, chairman of the corporate membership committee at the American Institute of Aeronautics and Astronautics, during the House subcommittee hearing. "We simply do not know what is next," he said. ■

BIOMEDICAL RESEARCH

Rare-disease project has global ambitions

Consortium aims for hundreds of new therapies by 2020.

BY ALISON ABBOTT

Prader–Willi syndrome. Fabry renal disease. Spinocerebellar ataxia. Few people have heard of these and the other 'rare diseases', some of which affect only hundreds of patients worldwide. Drug companies searching for the next blockbuster pay them little attention. But the diseases are usually incurable — and there are thousands of them.

This week, the US National Institutes of Health (NIH) and the European Commission launch a joint assault on these conditions, whose small numbers of patients make it difficult to test new treatments and develop diagnostic methods. The International Rare Disease Research Consortium being formed under the auspices of the two bodies has the ambitious goal of developing a diagnostic tool for every known rare disease by 2020, along with new therapies to treat 200 of them. "The number of individuals with a particular rare disease is so small that we need to be able to pool information from patients in as many countries as possible," says Ruxandra Draghia-Akli, the commission's director of health research.

At the launch meeting in Bethesda, Maryland, on 6–8 April, prospective partners will map out research strategies to identify diagnostic biomarkers, design clinical trials and coordinate genome sequencing in these diseases. Nearly all the rare diseases, of which there are an estimated 6,000–8,000, are the result of small genetic changes.

The meeting will also discuss the governance of the project, which is most likely to be modelled on the pioneering Human Genome Project. As such, the consortium is open to research agencies and organizations from all over the world. Representatives from countries including Canada, Japan and some individual European nations are all attending the meeting, and may join the consortium. Those wishing to participate will have to pledge a minimum financial contribution, which has not yet been agreed, and share all relevant data. Indeed, the project will have to overcome numerous obstacles to information sharing, such as the fact that physicians in different countries often use entirely different words to describe the same disease.

Draghia-Akli points out that the project could yield major benefits for the emerging field of personalized medicine — another political priority for the NIH and the commission — which also faces the challenge of small populations of patients.

"We need to be able to pool information from patients in as many countries as possible."

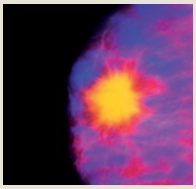
Regulatory agencies such as the US Food and Drug Administration and the European Medicines Agency rely on large, randomized and controlled clinical trials when deciding whether to approve new medicines, and one of the aims of the consortium will be to develop alternative clinical-trial methods for diseases that affect few people.

These methods are becoming ever more important now that genome analysis is helping to break down common diseases into ever smaller subclasses. "Soon there will be no disease called breast cancer," says Draghia-Akli. Instead, the catch-all term will be replaced by "a large number of rare diseases, each of which causes malignant growth in breast tissue and requires individual treatment", she says.

The commission will launch a €100-million (US$140-million) call for research proposals in July, which will support the consortium's scientific goals by focusing heavily on developing appropriate clinical trials. ■

MORE ONLINE

TOP NEWS



KINGS COLLEGE SCHOOL MED./SPL

Fifty genome sequences reveal breast cancer's complexity go.nature.com/7jfuji

MORE NEWS

- Carbon-rich mangrove forests are ripe for conservation go.nature.com/1djphk
- The amazing disappearing antineutrino go.nature.com/cvm8xi
- China vows to clean up rural environment go.nature.com/extggt

FUKUSHIMA EXPERT ANALYSIS

Andrew Sherry: The technology that makes modern nuclear reactors safer than Fukushima go.nature.com/9gbk9c

David Brenner: Why we still don't understand the risks of low doses of radiation go.nature.com/67ygor

© 2011 Macmillan Publishers Limited. All rights reserved

FUNDING

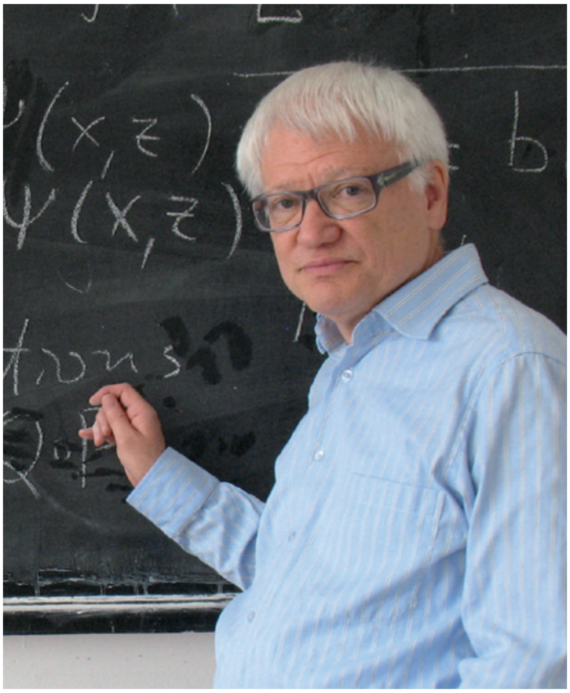
Bulgarian funding agency accused of poor practice

Former director claims his efforts to reform National Science Fund have been hindered.

BY ALISON ABBOTT

Science in Bulgaria, already hobbled by a cash shortage and by faltering reforms of communist-era institutions, is facing a new setback. Its main source of cash, the Bulgarian National Science Fund (NSF), has been accused of mismanagement — by its own former director, Emil Horozov. An investigation commissioned by Horozov found widespread irregularities in the NSF's handling of funding requests in 2008 and 2009, including using unqualified referees, and selectively ignoring referees' comments to favour particular projects.

When a summary of the investigation's report was published online in March, four months after the report was delivered to the country's Ministry for Education, Youth and Science, angry Bulgarian scientists accused the minister, Sergei Ignatov, of suppressing the findings.



Emil Horozov has quit as head of Bulgaria's main funding agency.

M. YANCHEV

Ignatov, who became minister in 2009, says that he had not suppressed the report, and had passed it to a committee of the finance ministry in December 2010 to review the claims of funding mismanagement. "Such serious allegations need professional analysis," he says. "I expect to receive a response in the very near future."

Horozov, a reputed mathematician at Sofia University, commissioned the investigation a few weeks after he became director of the NSF in January 2010. He had noted that many of the NSF's funding decisions seemed to be being made in a secretive way. So he appointed a working group of eight independent scientists to examine the refereeing and selection of the nearly 2,000 proposals competing for a share of 250 million leva (US$181,000) over the previous two years. The committee delivered a detailed report in October 2010, which was passed to the ministry the following month.

The report claims that many of the 230 reviewers who were chosen to assess the proposals had no science degrees or academic position. "One professor of English reviewed a paper in biochemistry," says Horozov. Each of the reviewers was allowed to select which grant proposals he or she would review from the whole list, a practice that is unheard of in funding agencies. As the reviewers were paid for each review they produced, Horozov believes that this system may have been abused by some reviewers to boost their fees. "Several referees, even those without any form of scientific qualification, managed to comment on 100–200 applications within just two months."

MISDIRECTION

In 2009, about 200 projects were selected for funding. In 82 cases, however, the investigation found that one or more of the reviews had not been taken into account; had they been, all but three projects would have fallen below the funding threshold. Although not illegal, this is certainly bad practice, says Horozov. About 50 reviews relating to unfunded projects were found to have been withheld. All of them gave higher scores than the projects' averages — had they been included, 12 of the projects would have been raised above the funding threshold. "If the allocation of money had been made randomly, the results would have been fairer," says Horozov, adding that, "Bulgaria is a poor country. If the very small amount of research money it does have had been spent well, it could have done a lot of good."

➲ NATURE.COM
For more on Bulgaria's science crisis, see go.nature.com/6icjlr.

The investigation also found that some of the referees reviewed competitors' applications, violating conflict-of-interest rules. Furthermore, some thematically unrelated, low-scoring proposals were pooled to give a higher overall score, which was then used to justify giving full funding to each individual project. The report also claims that the NSF paid a company the highly inflated price of 186,000 leva for an online grant-proposal submission system without putting out a tender, as the law requires.

After waiting for more than three months to get a response from the science ministry, Horozov resigned as NSF director on 23 February, saying that Ignatov had also hindered his efforts to prevent further corruption in the agency. Horozov says that when he became director, he understood that he would have the power to reform the NSF, modelling its structure on that used by successful European research-funding agencies. He presented a proposal on this to the minister in October 2010 but says he got no response. Ignatov says he received no concrete proposal, but that his ministry had in any case been busy with many other reforms.

Two weeks ago, Horozov posted a summary of the investigation on his webpage (see go.nature.com/exqfy5) and described its contents in open lectures at the University of Sofia and the Bulgarian Academy of Sciences. On 24 March the Civil Movement for Support of Science and Education in Bulgaria, a grass-roots organization of academics, issued a statement of support for Horozov, and called for Ignatov to resign.

Pavel Kerchev, a Bulgarian PhD student at the University of Leeds, UK, says that the scandal comes as no surprise. "Bulgaria is a small society and scientists have long been aware that the NSF was not working properly. It was just a matter of time before someone had the courage to bring it out publicly." ■

© 2011 Macmillan Publishers Limited. All rights reserved

GENETICS

Patent dispute threatens US Alzheimer's research

Lawsuit could expose hundreds of scientists to property-rights litigation.

BY ERIKA CHECK HAYDEN

The website of the Alzheimer's Institute of America (AIA) doesn't reveal much about the organization, but portrays it as committed to supporting research and patients. Among people who study Alzheimer's disease, however, the AIA, based in St Louis, Missouri, is best known for filing lawsuits against companies and researchers — a practice that scientists say could hamper the progress of research into combating the dreaded disease.

An AIA lawsuit filed in February 2010 against the Jackson Laboratory in Bar Harbor, Maine — a source of laboratory mice funded by the US National Institutes of Health (NIH) — now threatens hundreds of government-sponsored Alzheimer's researchers with litigation. The lab is so concerned about the financial and scientific costs of defending itself that it has asked the NIH to assume the defence of the case.

"The lawsuits raised by the AIA are unfortunate, and constitute a large drain on valuable scientific resources at a time when scientific funds are increasingly tight," says Benjamin Wolozin, an Alzheimer's researcher at Boston University in Massachusetts.

The suit concerns an AIA patent on a human DNA sequence used in mouse models of Alzheimer's disease. The sequence encodes the 'Swedish mutation' (discovered in a Swedish family), which causes early-onset Alzheimer's. Michael Mullan, a biomedical researcher who is now head of the Roskamp Institute in Sarasota, Florida, patented the sequence in 1995, then sold it to the AIA.

The NIH requires scientists to share transgenic mouse strains developed using NIH money, and the agency funds Jackson to breed, house and distribute these mouse models, says David Einhorn, house counsel at the lab. The AIA is alleging that Jackson infringed on its Swedish mutation patent, and others, when the lab distributed 22 strains of mice with the mutation to researchers; the organization is seeking unspecified damages.

The lawsuit also accuses six commercial companies of improperly profiting from the Swedish mutation, for instance by using mice bearing the mutation to test potential drugs. Furthermore, the AIA has filed four separate suits relating to the patent against academic institutions and companies in Oklahoma, Florida, Missouri and Pennsylvania (see 'Patent disputes').

But the litigation against Jackson could have the broadest impact on research. According to Einhorn, the AIA is demanding that Jackson hands over the names of all scientists who have worked with the relevant mouse models; this raises the possibility that those individual researchers might also be sued.

Last month, judge Elizabeth Laporte for the US District Court of Northern California recognized the potential impact of the suit on Alzheimer's research. She denied an AIA request to expand the suit by adding another patent-infringement claim, writing in her decision that the AIA has not disputed Jackson's claim that "prolonging the litigation in this case would harm Jackson and the public by extending the chilling effect of the litigation on mice research on Alzheimer's disease".

The AIA says that it allows academic research on mouse models covered by its patents, but does not permit work that profits from them. "Jackson Laboratory is not giving away the mice for academic research. On the contrary, these mice are being sold, and Jackson Laboratory is making quite a lot of money in the process. Furthermore, the mice Jackson sells are, in many instances, being used for commercial, not academic, purposes," the institute wrote in a statement.

Einhorn counters that the lab doesn't make enough from distributing mouse models to cover its operating costs, and it relies on philanthropy and public and private grants to support its work. He says that Jackson only allows academics, not companies, to use the models, and points out that asserting rights in such cases runs counter to common practice, which is established by NIH policy.

Defending against the lawsuit puts Jackson in a difficult spot. Proving that the AIA's allegations are groundless could take years and millions of dollars. It could also cast a pall over the Alzheimer's-research field, which has already been scarred by an extensive fight over the Swedish mutation patents during the 1990s (see *Nature* **404,** 319–320; 2000). But settling out of court would require Jackson to hand over researchers' names, a demand that Einhorn calls "repugnant".

"We haven't been able to settle this case because we're trying to do the right thing by trying to support the NIH policy and protect researchers out there in the community," says Einhorn.

Kathy Hudson, NIH deputy director for science, outreach and policy, says that the agency is considering the lab's request for help, made last December. "We're trying to evaluate the legal risks and the risks to the research community," she says. ■

J. L. TORRANCE/JACKSON LAB

Transgenic mouse models used in research are at the heart of the litigation.

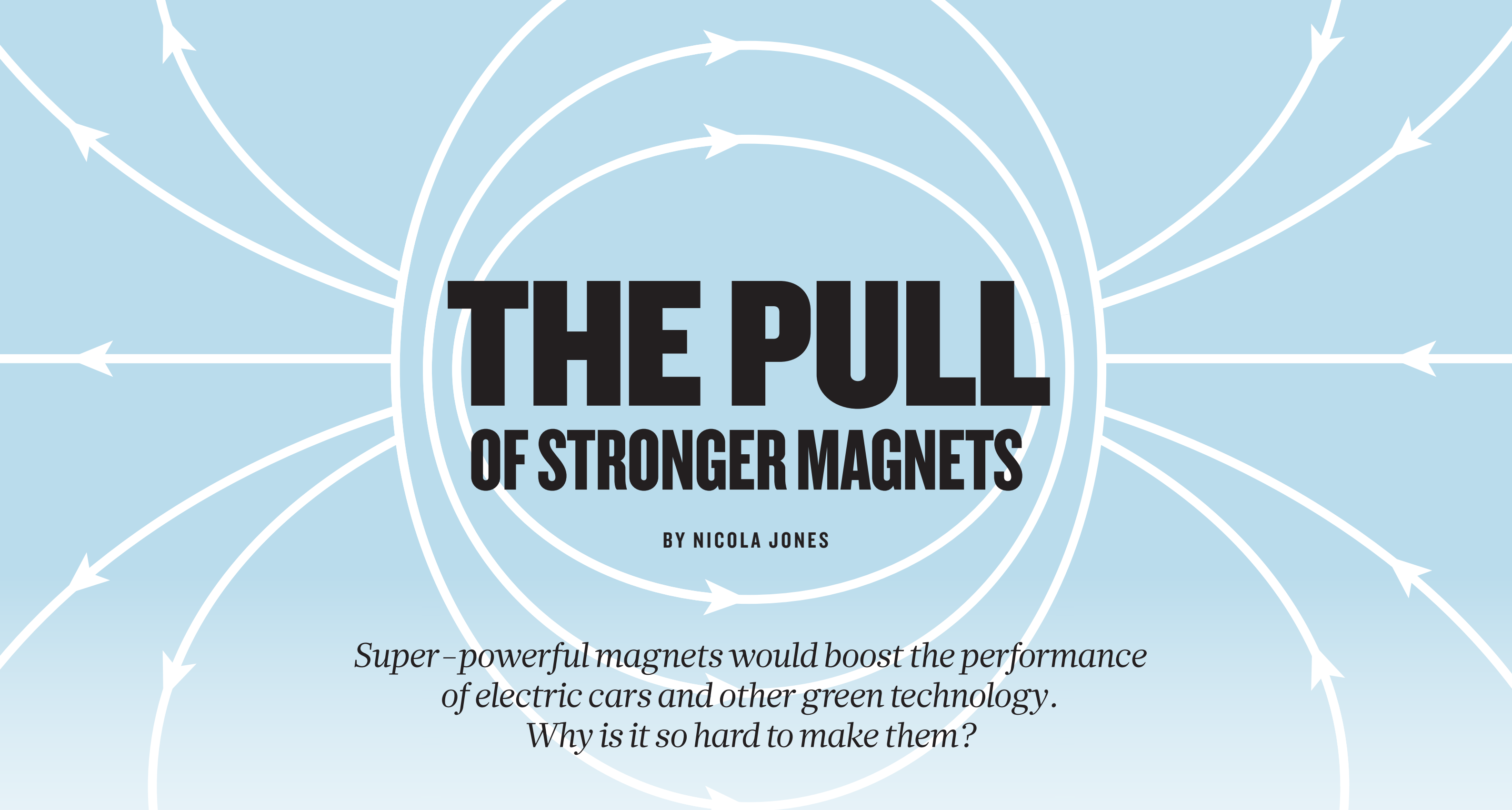
ONGOING LAWSUITS

Patent disputes

The Alzheimer's Institute of America (AIA) in St Louis, Missouri, is involved in multiple lawsuits regarding alleged infringement of its patents.

- AIA vs University of Pennsylvania and Avid Radiopharmaceuticals. Filed November 2010 in Pennsylvania.
- AIA vs Jackson Laboratory, Elan, Eli Lilly, Anaspec, Immuno-biological Laboratories, Invitrogen and Phoenix Pharmaceuticals. February 2010; Northern California.
- AIA vs Oklahoma Medical Research Foundation and Comentis. December 2009; Oklahoma.
- AIA vs Pfizer. June 2009; Missouri.
- Mayo Clinic Jacksonville vs AIA. March 2005; Florida.

© 2011 Macmillan Publishers Limited. All rights reserved



THE PULL OF STRONGER MAGNETS

BY NICOLA JONES

Super-powerful magnets would boost the performance of electric cars and other green technology. Why is it so hard to make them?

For Christmas, magnetics researcher William McCallum got one of the latest cool toys: 'Buckyballs: The Amazing Magnetic Desktoy You Can't Put Down!' The magnets are state-of-the-art — strong enough that, if they were cubes rather than spheres, you wouldn't be able to pry them apart. But if McCallum has his way, his team will make them look like weaklings.

McCallum, a materials scientist at Iowa State University in Ames, is tackling two big problems at the same time: magnet strength and cost. For most of the twentieth century, the strength of available magnets doubled every decade or two, but it stalled in the 1990s. The limit has hampered efforts to make high-tech products such as electric cars more efficient. And in the past two years, the cost of the rare-earth elements that are essential to advanced magnets has shot up. The price of neodymium oxide jumped from US$17 a kilogram to $85 a kilogram in 2010 alone.

Despite their name, rare-earth elements such as neodymium aren't truly rare geologically, but they are expensive to mine and process. China, which provides about 95% of the 96,000 tonnes currently produced worldwide every year, has put increasingly stringent caps on exports, even as the need for the elements is booming. Magnets made with them are at the heart of modern technology from mobile phones and laptops to high-efficiency washing machines. And many devices that are part of the green economy require substantial amounts: an electric car carries a few kilograms of rare-earth elements, and a 3-megawatt wind turbine uses about 1.5 tonnes. Demand leapt from 30,000 tonnes in the 1980s to 120,000 tonnes in 2010 (which was met in part by depletion of national stockpiles), and is predicted to hit 200,000 tonnes by 2015, says Gareth Hatch, founder of the Technology Metals Research consultancy in Carpentersville, Illinois (see 'Market forces').

Fortunately, the leading idea for how to make 'next-generation' magnets could solve both problems at once. It involves combining nanoparticles of rare-earth magnets with nanoparticles of cheaper magnetic materials — creating super-strong end-products with far less of the expensive ingredients. Governments keen to invest in energy-efficient technology, and scared by a global crunch in the rare-earth market, have started to pay attention to magnetics research.

In the United States, an infusion of funds has come from the Department of Energy, home of the Advanced Research Projects Agency — Energy (ARPA-E), which was established in 2009 to bring high-risk, potentially 'transformative' technologies to the market. ARPA-E has allocated $6.6 million to research on next-generation magnets — a shot in the arm for the field. "We're long overdue" for the next magnet revolution, says George Hadjipanayis, a physicist at the University of Delaware in Newark, who is head of a $4.4-million ARPA-E consortium of which McCallum is part. "We need to do it."

Permanent magnets get their pulling power from the orbits and spins of unpaired electrons, which tend to align with an external magnetic field and stay that way when that field is taken away. These magnets are ranked by their 'energy product' in kilojoules per cubic metre (kJ m^{-3}) — a combination of how much they respond to an applied magnetic field (their magnetization) and how well they resist being demagnetized. These properties don't always go hand in hand. Iron–cobalt alloy has the highest potential magnetization known, but its energy product is effectively zero because it is easily demagnetized: it has a symmetrical cubic crystal structure, with nothing to keep its electron spins pointing in any one direction, so they can be jolted out of alignment by a bump or a nearby magnetic field.

SPINS IN SYNC

Newer magnetic materials have a complex crystalline structure that helps to keep the spins pointing one way. In the 1950s, the best of such magnets, made of an alloy of iron, aluminium, nickel and cobalt called Alnico, achieved an energy product of 40 kJ m^{-3} (see 'Stalled progress'). The 1960s brought the first generation of rare-earth magnets, made of samarium and cobalt, which eventually enabled energy products to exceed 250 kJ m^{-3}. In the 1980s, researchers devised neodymium–iron–boron (NIB) magnets, which hold the record at about 470 kJ m^{-3}. If the magnets have to work at high temperatures — such as in a car engine — the rare-earth element dysprosium is added to the mix.

The dream is to unite the magnetic punch of something like iron–cobalt with the stability of, for example, a NIB magnet. That should be possible by combining nanoparticles of the two, packed so closely that neighbouring electrons influence each other and keep their spins aligned. In theory, a nanocomposite could reach an energy product of a whopping 960 kJ m^{-3}, with rare earths making up just 5% of its weight, compared with 27% in a normal NIB magnet (R. Skomski and J. M. D. Coey *Phys. Rev. B* **48,** 15812–15816; 1993). But making such a composite is extremely difficult.

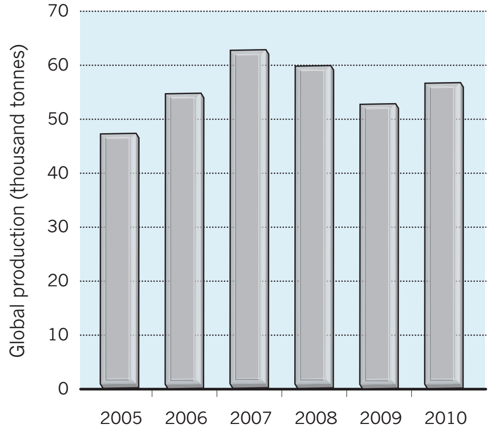
The grains in a successful nanocomposite must be small (10 nanometres or less); have the right crystal structure; have aligned magnetic directions; and be tightly packed. Achieving all of these at once is an engineering nightmare. On top of that, rare-earth nanoparticles aren't stable — they love to react with oxygen, which ruins their magnetic properties.

➲ NATURE.COM
Read more on the search for efficient energy technology at:
go.nature.com/jyfb2i

© 2011 Macmillan Publishers Limited. All rights reserved

MARKET FORCES

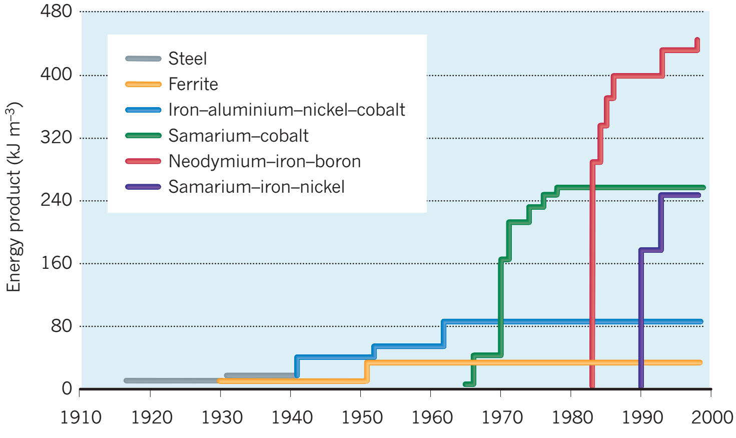
Sharply rising demand for neodymium–iron–boron magnets drove production up rapidly until 2007, but the global economic crisis caused a brief downturn.



SOURCE: W. T. BENECKI, T. K. CLAGETT & S. R. TROUT *PERMANENT MAGNETS 2010–2020* (2010)

STALLED PROGRESS

For most of the twentieth century, the strength of magnets jumped up every decade or so, with the introduction of new materials. The improvement has now slowed, but researchers hope to make the next leap soon.



SOURCE: O. GUTFLEISCH *ET AL. ADV. MATER.* **23,** 821–842 (2011)

In 2006, a team led by Ping Liu, a physicist at the University of Texas at Arlington, pioneered a manufacturing method that used steel balls to grind up magnetic material with the desired crystalline structure in a solution containing detergents. "I had postdocs working for years on this before we got a publication," says Liu. "They hated me." The soap lets the team produce nano-sized grains that don't adhere to each other but do keep their magnetic properties. Hadjipanayis is using the same technique, and says that in the past year he has made grains as small as 2.7 nanometres.

Even more difficult is making a bulk magnet out of these grains. One standard technique — pressing the grains together and heating them to 800–1,000 °C — causes them to diffuse into each other, so they become too big to create the cooperative nanocomposite effect. Another method — using polymer glues to bind the grains — dilutes the magnetic material.

There are alternatives. Hadjipanayis plans to charge one set of nanoparticles positively and the other negatively, so that electrostatic attraction binds them together. Liu's group squeezes about half a gram of the nano-grains in a press for 30 minutes instead of the standard half a minute. He also adds a bit of warmth (about 500 °C) to help them deform, but not so much as to ruin them. Using this method, Liu has managed to make relatively strong, dense magnets, but the grains aren't magnetically aligned, so the magnets are still weaker than a standard NIB one.

"YES, IT IS AMBITIOUS, BUT THAT'S EXACTLY WHY WE NEED TO BE DOING IT."

Alignment is the final hurdle. Liu's group is trying to clear it by putting material through a second slow- compaction process, but is having limited success. The researchers are fiddling with the details, trying to hit on a recipe that works. "I hope it can be done before my retirement," says Liu.

CORPORATE COMPETITION

Liu could be beaten by his competition before he reaches that deadline. The technology firm General Electric, headquartered in Fairfield, Connecticut, has been given a $2.2-million ARPA-E grant to pursue nanocomposites, and has beefed up its magnetics research team. The company, which started its experimental work in January, told *Nature* that it has a good way to make crystalline grains, but it wouldn't give details.

Last December, the US Department of Energy released its *Critical Materials Strategy*, which outlines a three-part mission to deal with shortages in rare-earth elements: secure new supplies, promote recycling and conduct research into alternatives, such as next-generation magnets. This push toward stronger magnets is a welcome change with potentially big pay-offs, says Liu. According to his calculations, doubling the strength of a magnet in an electric car should improve the motor's efficiency by about 70% — although that number could vary wildly depending on the design of the magnet and engine.

Although the United States seems to be making the most concerted push towards creating the strongest magnets, other nations have invested more money in general magnetics research, says Liu. China's 5-year economic plan for 2011–15 includes a big boost — reportedly more than 4 trillion renminbi (US$610 billion) — for spending in seven 'strategic emerging industries', including energy systems, clean cars and new materials. Observers such as Hatch and Liu expect great things from the investment. Japan has invested heavily in magnet research for its high-tech industry, and has strong government–industry collaborations — although one of its largest centres for magnetics research is Tohoku University in Sendai, which was hit hard by the earthquake and tsunami in March (see *Nature* **471,** 420; 2011).

Last year, the European Union's research-funding framework put out a €4-million (US$6.3-million) call for proposals from groups working to develop novel materials, with the goal of totally replacing rare earths. But most researchers say that this is massively overreaching. "This is a joke, scientifically," says Liu of the quest to remove rare earths from strong magnets. Several major labs have had proposals rejected because they aimed simply to reduce the quantities of rare earths used in magnets, says Dominique Givord, a magnetics researcher at the Louis Néel Laboratory in Grenoble, France.

Researchers' target of building next-generation nanocomposite magnets is, most admit, a long shot. "I know that this activity is becoming popular in the United States, but I feel that their goal is a bit too ambitious," says Kazuhiro Hono, a magnetics researcher at the National Institute for Materials Science in Tsukuba, Japan. Givord agrees. "It is extraordinarily challenging," he says. More realistic, he says, are attempts to make existing magnets a bit stronger and cheaper by altering their microstructures. In Japan, such efforts have helped to reduce dysprosium demand.

But Hatch, who has worked in the field for nearly two decades, says that next-generation magnets are worth the battle. "Yes, it is ambitious, but that's exactly why we need to be doing it," he says. "It's time to put money behind it." ■

Nicola Jones *is a freelance journalist based near Vancouver, Canada.*

© 2011 Macmillan Publishers Limited. All rights reserved

EDUCATING INDIA

The country's vast, education-hungry population could supply the next generation of the world's scientists — but only if it can teach them.

BY ANJALI NAYAR

Subha Chakraborty has hardly left the lab in three months. His master's research in micro-scale systems is running into the early hours almost every morning, and "that is not the right time to go back to your room and sleep", he says. So he bunks on a makeshift bed under his computer and cooks on a toaster in the corner of the lab's common room.

Chakraborty isn't alone: most of the lab's ten postgraduate students follow a similar schedule. "There's some kind of charm here," says one of them, Anindya Roy, who has decided to officially surrender his dormitory room.

These students at the banyan-tree-lined campus of the Indian Institute of Technology (IIT) in Kharagpur are among India's luckiest and best: once they have completed their degrees, they will end up working at top universities and private research hubs in India and around the world. But the optimism and drive are ubiquitous. "When you go to the rural parts of the country you meet extraordinarily bright kids who just have to be given the opportunity," says Chintamani Rao, chief scientific adviser to India's prime minister. There are a lot of them — around 90 million between the college-going ages of 17 and 21, rising to an estimated 150 million by 2025. And they are hungry, starving even, for an education.

BRAIN DRAIN

But can India feed that hunger? The government has pledged to make it a priority, but faces tremendous obstacles. Most of the elite science and engineering graduates opt for high-paying jobs in industry rather than independent research. Other students far too often end up in high-priced commercial diploma-mills that deliver little real education. Many, many more young Indians don't even get that far: the country's 500 universities and 26,000 colleges have space for only about 12% of its eligible youth. And the population is growing by 1.34% a year, more than twice the rate of growth in China (see 'A double explosion').

But if India cannot meet this challenge, it could miss out on becoming one of the world's great innovation hubs, says Rao. "There is a very large population out there that is extremely qualified and they end up in second or third-rate institutions," agrees Pradeep Khosla, dean of engineering at Carnegie Mellon University in Pittsburgh, Pennsylvania, and a graduate of IIT Kharagpur. "A lot of talent gets wasted."

On the surface, India seems to be in the middle of an educational renaissance, thanks largely to its booming economy. After decades of economic stagnation under the socialist policies that followed the country's independence in 1947, Indians enthusiastically embraced a series of business-friendly reforms that began in the early 1990s. The result has been economic growth that currently averages more than 8% a year, with only a slight and temporary slowdown during the global financial crisis that began in 2008. That growth, in turn, has created a flourishing market for qualified graduates in everything from construction to information technology and health care.

"There are a lot of stories of successes — from rags to riches — of Indians who made it just on the basis of good education," says Pawan Agarwal, author of *Indian Higher Education: Envisioning the Future* (Sage; 2009). "This is creating high aspirations among Indians about higher education."

Those ambitions, along with the population growth, have fuelled an eight-fold increase in science and engineering enrolment at India's colleges and universities over the past decade, with most of the growth

© 2011 Macmillan Publishers Limited. All rights reserved

L TO R: A. NAYAR; PHOTOLIBRARY.COM; B. CRUICKSHANK/LONELY PLANET IMAGES; DINODIA

occurring in engineering and technology — fields in which jobs are especially plentiful. The low cost of doing business in India and the large crop of English-speaking graduates has made it a global hot spot for investment in research and development (R&D).

"In 2003, 100 foreign companies had established R&D facilities in India," says Thirumalachari Ramasami, head of the government's Department of Science and Technology. "By 2009, the number had grown to 750." Those companies include technology and communications firms such as IBM, General Electric, Cisco, Motorola, Oracle and Hewlett-Packard, all eager to get a foothold in the fast-growing information-technology hub around Bangalore.

Small wonder, then, that the 15 IIT campuses nationwide have roughly 300,000 applicants every year, or that the students who make it in are very, very good: IIT acceptance rates are about 2% (see 'Only the best'), compared with around 7% at Harvard University in Cambridge, Massachusetts, an emblem of US elitism. "Statistically, out of a billion people there must be a Michael Faraday," says Rao. "There must be a number of talented people."

Look closer, however, and it becomes apparent that there are serious cracks in the system. For example, the vast majority of India's science and technology graduates immediately head for high-paying jobs in industry. Only about 1% of them go on to get PhDs, compared with about 8% in the United States. "Internally the brain drain is quite high," says Rao. "All the talent goes into sectors that make money but produce very little in terms of creative things for the country."

What makes this problematic, adds Rao, is that the country's rising economic tide is largely the result of its myriad outsourcing centres and the computer industry. If India cannot broaden its economy — and make better use of its brightest scientific minds — it will have little chance of solving its challenges in areas such as poverty, food, energy and water security.

➲ NATURE.COM
How technology is helping India's children to learn:
go.nature.com/iehycd

"Everyone's just making computers faster, and our computers are pretty fast already," agrees Manu Prakash, who graduated from the IIT in Kanpur — and who, like many Indians with academic ambitions, elected to pursue his education elsewhere. He earned his PhD from the Massachusetts Institute of Technology in Cambridge, and now runs his own biophysics lab at Stanford University in California.

Prakash says that although the IIT system does attract superb students, it is institutionally broken because it doesn't value creativity. "You have a brilliant mathematician coming into an engineering course and then taking a nine-to-five job with a company," he says. "There is something wrong there."

QUANTITY VERSUS QUALITY

Whatever its flaws, the IITs remain out of reach for millions of eager, ambitious Indian students. The higher-education system is expanding pellmell to accommodate them — with the burgeoning private sector filling around 90% of the demand. "We will need another 800–900 universities and 40,000–45,000 colleges within the next 10 years," says Kapil Sibal, India's minister of human resources and development. "And that's not something the government can do on its own."

"WE ARE SPOON-FED. THE TEACHERS DICTATE AND THE STUDENTS WRITE DOWN WHAT THEY SAY."

For-profit colleges and universities are popping up around the country by the day — nearly 4,000 of them in 2010 alone. The road leading out of Chennai in southern India, like many around the country, is crammed with hundreds of private engineering colleges. The government has struggled to maintain any kind of standard. "The big challenge is that when you move to grant more access [to education], that the access must come with quality," says Sibal.

Many private institutions have only a few hundred students each and offer little in the way of laboratory or practical training, because labs are expensive. Curricula are outdated and there are crippling shortages of teaching staff, thanks to the allure of higher-paying industry jobs. "The younger generation is completely disillusioned with pursuing higher education with the intention of going into teaching," says Agarwal. Sibal estimates that at least 25% of academic posts are vacant and more than half of professors lack a postgraduate education.

Rahul, who prefers that his real name not be used, studies information technology at a private college an hour outside Delhi. "We are spoon-fed," he says. "The teachers dictate and the students literally write down what they say."

Rahul's parents paid hundreds of thousands of rupees up front to get him into the institute after he scored poorly on entrance exams. He says that about 30% of his peers entered in the same way, and at other colleges the informal 'management quota' can be as high as 40–50%.

This year, tuition at the institute cost 85,000 rupees (US$1,900): more than three times that charged by the IIT system. And the payments at many private colleges don't stop there, says Rahul. "A few days before [exams] you can pay 1,000 rupees for a copy of the paper, and you can pay another couple of thousand rupees if you didn't get the right marks," he says. "Then, if you don't attend classes or labs, you can pay 5,000 rupees to fulfil your attendance quota. Education here is based entirely on money. And to think, my institute is one of the best in the area."

There are more than 600 colleges affiliated with one university in his province alone, and every college has 5–6 branches, with 60–120 students each. "That's lakhs [hundreds of thousands] of students passing out of these colleges per year," says Rahul.

Moreover, many of the students are graduating with abysmal literacy and numeracy skills.

© 2011 Macmillan Publishers Limited. All rights reserved

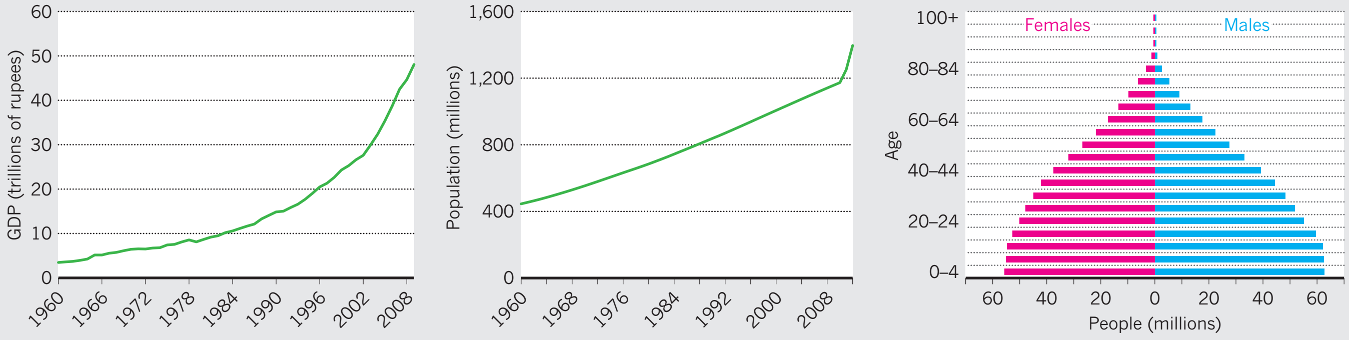
A DOUBLE EXPLOSION

India is struggling to meet rising aspirations for education.

India's booming economy is spawning millions of new jobs – especially for college-educated scientists.

But India's population is expanding even faster, topping 1.2 billion last month.

The result is a country full of young people, many more than India's educational system can accommodate.



Employers' surveys suggest that up to 75% are unemployable.

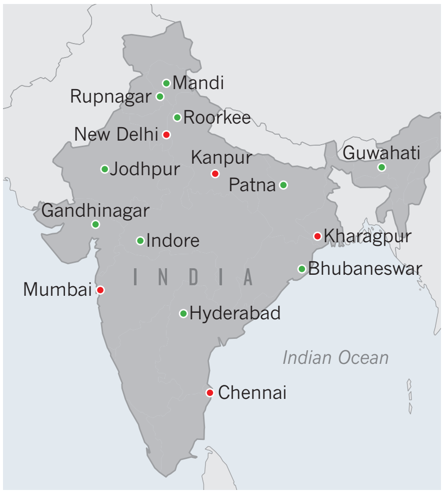
"You can pay to get in, you can pay to get good marks and you can pay for your attendance, but you can't pay to get into a good company," says Rahul. "There are people at my college who don't even know how to say 'how are you?' in English" — the working language of most companies.

Rahul's experience is not unusual. Geeta Kingdon, who studies education, economics and international development at the University of London's Institute of Education, points to allegations of widespread corruption in how Indian institutes and universities are accredited. "Even those who have got the relevant accreditation only got it because they paid the relevant bribe," she says. Many don't bother. A government crackdown on unaccredited institutions in 2010 left more than 40 universities and thousands of colleges in court.

Corruption has even reached the august halls of IIT Kharagpur. Last October, a handful of the institute's top engineering professors were accused of running a fake college called the Institution of Electrical Engineers (India) from the campus. The scheme allegedly involved the use of forged documents bearing the IIT logo to lure in students, who were charged 27,000 rupees for admission, roughly what the IITs charge per year. The IIT Kharagpur has launched an inquiry into the incident. "But there will always be another scandal down the road," says Srinivasan Ramanujam, a mechanical engineer at the institute. "Students are desperate to get into a college and people exploit this mentality."

ONLY THE BEST

Even with ten new campuses established since 2000 (green), the Indian Institute of Technology system accepts only around 2% of the 300,000 who apply.



With all these desperate but half-baked graduates, India's hopes of becoming a global centre of innovation are being compromised. Too often, the corporate R&D model sweeping through India treats science graduates more as grunt workers than true innovators, says Ramasami. "Just availability of scientifically talented people does not provide scientific breakthroughs. For the discovery process you need ambience and creative people."

India's government is working hard to change the trend. In January 2010, for example, it pledged to ramp its investment in R&D up from the current 1% of the gross domestic product to 2%, but this will happen very slowly, says Rao. The government's budget for 2011–12 included a one-third increase in its annual higher-education investment, to a total of 130 billion rupees. And it has approved a new funding agency, the National Science and Engineering Research Board, which is expected to become operational this year, and will have an initial budget of around US$120 million, says Rao.

By 2014, says Ramasami, the hope is that such measures will raise the number of science and technology PhDs awarded each year from the current 8,900 — less than one-third that of the United States or China — to at least 10,000. By the end of the decade, he says, the target is 20,000 PhDs a year.

OVERSEAS INPUT

The government is also counting on an injection of money and expertise from foreign academic institutions. With enrolment rates waning abroad, many universities are looking to India as a new academic market — including US institutions such as the University of California, Berkeley, and Carnegie Mellon University.

US President Barack Obama's trip to India last November highlighted the growing interest: included in his delegation were three presidents of US universities and senior representatives of several more. During the trip, Obama and Indian Prime Minister Manmohan Singh announced that they would hold a US–India summit on higher education this year to help encourage collaborations.

So far, Indian law has restricted foreign universities to forming partnerships with Indian institutions, says Sibal. But a Foreign Educational Institutions Bill being considered in India's parliament would allow them to build full-blown campuses of their own. Sibal takes it as a sign of what India could become. "Top-quality institutions of the United States and around the world are actually knocking at our door," he says. "The India of tomorrow will be an India that provides solutions not just for itself, but also for the rest of the world."

But that is only if India's rising youthful generation can break out of its current job-based mentality — not easy in a developing country.

One evening late last year, Shirsesh Bhaduri, a fourth-year biotechnology student at IIT Kharagpur, visited Tikka — a makeshift café in the shade of a banyan tree, where students and faculty members catch up over cups of 3-rupee tea and samosas. But just over the campus's whitewashed walls is the reality of West Bengal state and most of India: unruly fields, shanty villages, water buffalo and jungle.

"In other countries, people may choose their career according to their interests," says Bhaduri, who has just been to an interview with London-based bank Barclays. "But here the industries that pay the maximum attract the maximum applications. Most people do a master's in business administration after the IIT — and that is the aim of most people out here. Everything is money-oriented." ■

Anjali Nayar *is a freelance writer based in Nairobi.*

© 2011 Macmillan Publishers Limited. All rights reserved

COMMENT

NASA

Launch of the first shuttle, *Columbia*, in 1981. Will NASA have a future to rival its past?

NASA: what now?

This month marks 50 years since Yuri Gagarin first ventured into space in the Vostok 1 mission, and 30 years since NASA's first shuttle flight. As the shuttle *Endeavour* prepares for its final flight, seven experts outline what NASA's priorities need to be.

DENNIS BUSHNELL

Revolutionize research

Chief scientist at NASA Langley Research Center

To achieve revolutionary goals, such as sending humans to explore the Solar System, NASA needs to develop revolutionary technologies. Because it is extremely difficult to pick winners in advance, research and development is required in several areas simultaneously.

Transporting humans to low-Earth orbit — such as the International Space Station — and beyond are two very different missions. But they both depend on the same metrics: safety and cost.

Humans have been travelling to and from low-Earth orbit for some 50 years, mostly on what were once military rockets. Today's commercial rockets use similar expendable technologies. The development firm Space Exploration Technologies (SpaceX) in Hawthorne, California, achieved notable success last year with rocket launches at quite low cost. Thus, with due attention to safety, commercial transportation of humans to low-Earth orbit should be feasible.

Transport of humans outside low-Earth orbit, especially to the Moon, Mars and beyond, is a wholly different challenge. Aside from Apollo, which 'only' went to the Moon, we have almost no experience to draw on. Also, such expeditions become exceedingly costly with existing rocket technology if they are to guarantee that crew members will remain healthy during long missions.

Revolutionary technologies should be targeted at: reducing the mass of the vehicle; novel launch and propulsion systems (including alternative fuels, such as positrons, energy beaming and in-orbit refuelling); and intelligent architecture and systems for more affordable life-support and radiation protection. Several of these technologies could be truly game-changing. The use of nanotubes in spacecraft construction, for example, could reduce the 'dry mass' — the amount to be launched, excluding

▶

© 2011 Macmillan Publishers Limited. All rights reserved

fuel — by three to five times, if we can create structural materials with the same strength properties as individual nanotubes.

A final alternative to sending humans to the toxic environment of Mars would be to develop space exploration for everyone using immersive virtual reality and remote planetary sensors, with autonomous robotics to supply the data. This could offer a better-than-being-there experience at much reduced cost and risk.

MARC GARNEAU
Get us to Mars

First Canadian in space and now a Member of Parliament

I believe there is a specific challenge that can galvanize us all: sending humans to Mars. A clearly defined objective can seize the imagination. Neil Armstrong understood this when he criticized his country's decision to take a broader approach to space exploration rather than giving itself something with a specific end point, if not an end date.

Last year, President Barack Obama cancelled the Constellation programme intended to return astronauts to the Moon. As a result, the road map for human space exploration is no longer as clear as it was. NASA and other space agencies are about daring and inspiration. But while their engineers and scientists develop new technologies and make new discoveries, it is the public who must be mobilized to support human spaceflight. This happened with Apollo when there was a race to win, and money was no object then.

Today, there is no clear race to win and money is very much a limiting factor — but that doesn't mean there is no reason to once again attempt what seems impossible. For me, that should be an international human mission to Mars led by the United States. With the completion of the International Space Station, we have proved that many countries can work together and share both the cost and the development of new technologies.

JOHN M. LOGSDON
Build a case for humans in space

Professor emeritus at George Washington University

NASA will probably continue to muddle along once the shuttle retires. Over the past 20 years the United States has spent more than $20 billion on developing an alternative way to take humans into space. None even reached the flight-test stage. The 2003 *Columbia Accident Investigation Board Report* (of which I was an author) called the lack of a replacement for the shuttle "a failure of national leadership". That failure continues.

Eight years later, there is still no replacement in sight, just the hope that together, the private sector and NASA can develop ways to carry astronauts to the International Space Station (ISS) and replace NASA's embarrassing dependence on Russian rockets. Since the Columbia accident, NASA has used expendable vehicles to launch its science missions, so the end of the shuttle programme will have little impact on space science.

The biggest uncertainty is whether the United States will even have a human-spaceflight programme once the ISS is retired in 2020. In 2009, the Augustine Commission called for a spaceflight programme that is "worthy of a great nation". In the commission's view, that meant human exploration at increasingly greater distances from Earth. Since then, there has been a confused and confusing debate among the White House, US Congress, NASA and the non-government space community over the best way to get started. No compelling proposal has emerged. The case has not yet been made for going back to the Moon, visiting a near-Earth asteroid or sending humans to Mars. Until it is, the US leadership is unlikely to commit the country to human spaceflight "worthy of a great nation".

ROALD SAGDEEV
Send more robots

Former director of the Russian Space Research Institute and adviser to former President Mikhail Gorbachev

The closing down of NASA's space-shuttle programme leaves the Russian Soyuz rockets as the only spacecraft capable of delivering manned vehicles to the International Space Station (ISS). With the right political will, however, there is no reason why NASA cannot regain self-sufficiency in the next few years, even on a more modest budget.

In the interim, NASA has a genuine historic opportunity to rethink its goals once the ISS discontinues operations. An earlier vision to return astronauts to the Moon is off the agenda of the administration of President Barack Obama (and perhaps for the foreseeable future). A mission to Mars, a dream of spaceflight pioneers, in an environment of global multidimensional (not simply economic) crisis, will probably remain a dream for decades to come.

At the same time, the unmanned space programme is developing with tremendous success and is costing much less. Robotic missions have vastly enriched our knowledge of the Solar System, and of Earth in particular, and have put numerous new-generation telescopes into space. These developments challenge the need for a costly human presence in space. Yes, astronomers are thankful to NASA's shuttle astronauts for prolonging the life of the Hubble telescope, thereby making it so successful. And in its last flight on 19 April, the *Endeavour* shuttle will deliver to the ISS an alpha-magnetic spectrometer — the most advanced high-energy experiment yet to be deployed in space.

But if such ventures remain isolated episodes, the expensive game of human spaceflight risks degenerating into 'space tourism' paid for by taxpayers.

ED LU
Deflect risky asteroids

Physicist, entrepreneur and former shuttle astronaut

The reason for human spaceflight is to protect human civilization. That means preventing direct threats such as asteroid impacts on Earth, as well as opening up the Solar System to human activity, including commerce, science, exploration and, some day, settlement.

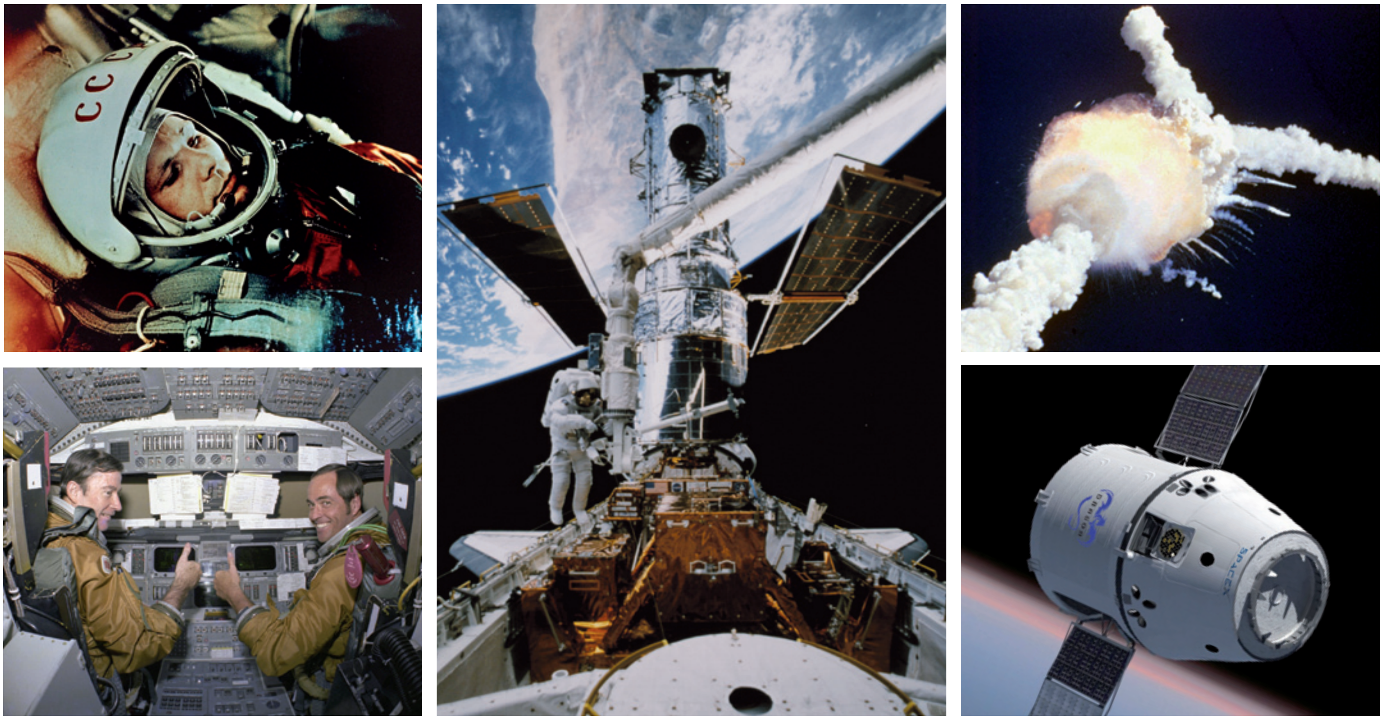
NASA should survey and catalogue the orbits of potentially threatening asteroids, and show that humans can alter the Solar System (if ever so slightly) by deflecting a non-threatening asteroid using a robotic spacecraft. Such a focus would tie together the human-spaceflight programme with the robotic planetary-exploration programme in a common purpose.

Most importantly, NASA must move faster. The agency moved so slowly on some recent major programmes that they have been cancelled for lack of progress. Its plans for a heavy-lift rocket should therefore also be scrapped: it is too expensive, and meaningful progress will not be made until the 2020s.

Instead, NASA should find ways to solve the fuel problem: the fact that most of the mass of any spacecraft leaving Earth is taken up by fuel. NASA should develop a fuel depot in low-Earth orbit that can be used to refuel missions to deep space. Commercial firms could be paid to deliver fuel to the depot. Routine operations of this sort will bring down the overall costs and free NASA up to develop its deep-space missions. The agency would then be able to make progress in extending the reach of humanity into the Solar System.

The author declares competing financial interests: details accompany the full-text HTML version of the paper at go.nature.com/wx4vlv.

© 2011 Macmillan Publishers Limited. All rights reserved



The highs and lows of human spaceflight. In 1961, Yuri Gagarin was the first man in space (top left). The shuttle first took flight 30 years ago (bottom left). Shuttle astronauts helped repair the Hubble Space Telescope in the 1990s (middle). The *Columbia* shuttle exploded after launch in 2003 (top right). Last year, a private company successfully launched a reusable space capsule (bottom right).

ROLLS PRESS/POPPERFOTO/GETTY IMAGES
NASA
MIDDLE: NASA. TOP: AP PHOTO
SPACEX

MATT MOUNTAIN
Find a united purpose

Director of NASA's Space Telescope Science Institute and scientist at the James Webb Space Telescope

Human spaceflight will be at its best when NASA can demonstrate that the whole is greater than the sum of its individual parts.

In the 1990s, NASA's ambitious shuttle missions to repair and upgrade the Hubble Space Telescope ensured that Hubble remained the most scientifically productive telescope in history, and uniquely captured the public's imagination. What stood these missions apart from other NASA human-spaceflight activities was that the whole agency was committed to a coherent purpose — a partnership between science and human spaceflight to explore the Universe — something that only NASA has done.

Now imagine a NASA committed to lead an international spaceflight endeavour to search for habitable worlds, and to extend humanity's reach to Mars and beyond — I suspect there would be a collective sigh of relief among the world's space agencies. It would give immediate focus and relevance to the International Space Station as a platform for understanding how to sustain a long-term human presence in space. For Mars, the focus should be not on rockets that hark back to the Apollo era — but on developing truly novel propulsion systems that allow humans to explore the entire Solar System.

In the medium term, NASA's astronauts could help to assemble and service giant space telescopes capable of searching for life around another star. The discovery of extra-terrestrial life would have as profound an impact on the twenty-first century as Neil Armstrong's Moon walk had on the twentieth.

NEAL STEPHENSON
Ditch the rockets

Science-fiction author and space enthusiast

NASA should throw itself into developing radically cheaper ways of getting into space: a task that only it can do, and that would help to restore the lustre and *esprit de corps* of a legendary organization.

Rockets got as good as they are ever going to get four decades ago. Measured in terms of specific impulse — the momentum imparted to the vehicle per unit of fuel, and the only factor that matters as far as the laws of physics are concerned — no game-changing advances have been made since the Apollo programme. The technologies pioneered by the Soviet Union and the United States have been endlessly cannibalized by NASA and parroted by many other countries.

The only way to fundamentally change humanity's relationship with space is to develop radically new launch systems, a challenge that no private company is likely to undertake. This is a job for NASA if ever there was one. The only catch is that it has to be NASA at its best — the NASA that many of us idolized in our youth — and not the grab-bag of aerospace-industry support programmes that the agency has become in the decades since the last Moon landings.

Scientists and engineers have been proposing alternative launch technologies since the 1950s, including laser- and microwave-powered propulsion, large gun-like devices, orbital tethers, space elevators, airplane- and balloon-assisted mechanisms and scramjets. None of these has taken hold, not because they are crazy (although some might be) but because the unbelievable amounts of taxpayers' money collected during the cold war and ploughed into old-school launch systems gave rockets a technological lead, and a privileged legal, regulatory and political position, unassailable by mere free enterprise.

Budget shortfalls provide an opportunity for NASA to eliminate many programmes that in happier economic times would be politically untouchable. NASA should make the most of this opportunity, and then rededicate itself to striving for the sorts of radical advances that, 50 years ago, had the power to awe the world.

© 2011 Macmillan Publishers Limited. All rights reserved

STOP THE MIRV
Universities are for students
Cities are for people
Science is for life
I-LABS MORE PRECIOUS THAN FAMILIES
WORKERS AND STUDENTS FOR ...PEOPLE

Protests against military research on MIT's campus flared on Alumni day on 16 June 1969.

MIT MUSEUM

The search for clean cash

The Massachusetts Institute of Technology has been an innovator of university funding models, says **David Kaiser**. Its 150-year history holds lessons for today.

One hundred and fifty years ago this week, on 10 April 1861, the Massachusetts Institute of Technology (MIT) received its charter. Although hardly the oldest institution of higher learning in the Anglo-American world — Harvard University was already well into its third century by then, and the British universities of Cambridge and Oxford were each on the cusp of their eighth — MIT quickly became a trendsetter. Founder William Barton Rogers built a curriculum around the school's motto *Mens et manus*: mind and hand. He and his faculty members incorporated laboratory instruction into the most elementary undergraduate courses and fostered close ties between basic science and the practical arts — pedagogical innovations that quickly inspired many imitators.

Perhaps the most influential of MIT's many innovations lay not in curricula or textbooks but in patronage. Time and again over its history, the institute has experimented with new ways to fund its research and teaching. While preparing a book on MIT, *Becoming MIT: Moments of Decision* (MIT Press, 2010), I learned just how vigorously the funding pendulum has swung between government and private funds. Every few decades, a decision inspired impassioned charges about whose money seemed appropriate or tainted, eliciting much hand-wringing from faculty members, administrators and alumni about what consequences might befall the scholarly community should the wrong choice be made. Each new scheme unleashed a battle for the soul of MIT. Yet proposals that had struck observers as bizarre or brash on first hearing were quickly absorbed into daily operations, and promptly emulated elsewhere.

Amid today's economic uncertainty, universities around the world again face difficult questions about how to fund their operations. MIT's experiences throw these struggles into sharper relief, showing that today's bandits were yesterday's heroes.

EARLY TRADE-OFFS

Just two days after MIT's charter was signed, mortar rounds began to fall on Fort Sumter near Charleston, South Carolina: the US civil war had begun. Although it hardly seemed a propitious start for the young institute, the outbreak of war bought Rogers time to continue searching for funds.

Fifteen months into the fighting, President Abraham Lincoln signed into law the Morrill Act. The law allowed individual states to sell federal land and use the profit to fund colleges that focused on applied or practical topics, such as agriculture or engineering. 'Land-grant colleges' quickly sprouted across the United States, nearly all of them public institutions. Rogers convinced the state legislature of Massachusetts to donate a handsome portion of its land-grant funds to the fledgling MIT — a private institution — and, in exchange, he promised to offer military instruction to all its students. The infusion of government cash convinced private donors that MIT was worth the investment. From the start, MIT thus functioned as a financial oddity: a private university buoyed by public funds.

By the end of the First World War, having fended off several merger attempts from nearby Harvard — which struck faculty members and alumni as hostile-takeover bids — MIT found its budget strained to the limit. In 1919, its president Richard Maclaurin launched a new campaign known as the Tech Plan. Until that time, MIT, like almost all other US universities, had relied on student tuition, private philanthropy and occasional grants from local industries to fund research.

Unlike those earlier efforts, the Tech Plan rebuilt MIT's entire operation around corporate patronage. It created a centralized

© 2011 Macmillan Publishers Limited. All rights reserved

Division of Industrial Cooperation and Research — the forerunner of today's ubiquitous technology-transfer offices — to facilitate corporate-funded research projects on campus, open the institute's libraries to industrial sponsors and share alumni records with corporate recruiters. Nothing like this had been attempted before in US higher education. MIT's Tech Plan immediately attracted hundreds of companies and generated hundreds of thousands of dollars (several million dollars in today's currency). From the start, it also courted controversy.

A decade after the Tech Plan was established, well over one-third of faculty members were conducting work for a corporate sponsor. What sort of work? Faculty members found it difficult to say, because many of their arrangements forbade publication of results without the sponsor's approval. Sentiment on campus for the Tech Plan further soured after the stock-market crash of 1929 and the onset of the Depression. Annual budgets for departments such as electrical engineering plummeted by 60% in just four years. A growing chorus concluded that MIT had been short-sighted to rely so heavily on corporate patronage which, after all, could be as fickle as the latest business cycle. Critics went further: overreliance on industrial funding was corrupting. In pursuit of quick money, the critics charged, MIT had auctioned off its intellectual autonomy.

GOVERNMENT'S TURN

Few alternative funding models seemed obvious. The Morrill Act notwithstanding, many fiscally conservative administrators at private universities across the country — from MIT to Stanford University in California — believed that the federal government had no business meddling in local affairs such as higher education. Self-made entrepreneurs, industrialists and philanthropists were one thing; federal bureaucrats were quite another. Their peers at public universities, who relied on state legislatures for funding, largely agreed. Yet the stark economic realities of the 1930s forced MIT vice-president Vannevar Bush to reconsider federal patronage. No industrial partners could rival the research budgets of projects such as the Tennessee Valley Authority, a sprawling, government-owned company founded in 1933 to investigate everything from agricultural productivity to hydroelectric power (see E. Rauchway *Nature* **457,** 959–960; 2009).

Bush's solution was to rely on contracts with the federal government. Both this funding source and the legal arrangement — contracts rather than grants or donations — were new. To keep up the appearance of fair-market transactions between autonomous agents, Bush insisted that MIT, desperate for cash though it may be, hammer out contracts at the negotiating table like any other private enterprise, rather than seemingly beg for handouts. Bush took this new approach with him to the brand-new National Defense Research Committee in June 1940, and its successor, the Office of Scientific Research and Development (OSRD), in 1941. On Bush's watch, the OSRD awarded thousands of research contracts to universities across the United States, for everything from radar to the atomic bomb.

More of those wartime research contracts flowed to MIT than to any other university. By 1945, MIT had secured defence-related research contracts worth three times more than those of stalwart industrial contractors Western Electric (AT&T), General Electric, RCA, DuPont and Westinghouse combined. More than 90% of MIT's annual operating budget derived from federal research contracts. In short order, MIT's model became the basic template for research universities across the United States. Other institutions (most famously Stanford) studied MIT's transformation and sought to replicate it. After the tremendous upheavals of wartime, few questioned whether the federal government was an appropriate source of funding for basic research and university education. The sheer scale of funding, which continued to rise after the onset of the cold war, quickly cemented the new norm.

During the 1950s and 1960s, federal patronage drove the fastest expansion of higher education in American history (if not the world). The new contracts, largely from military and defence-related agencies, underwrote massive new equipment on campus such as nuclear reactors and electronic computers, opening up unparalleled opportunities for faculty members and students. Few questioned the relationship too sharply until the escalation of fighting in the Vietnam War in the late 1960s. Only then did a critical mass of campus voices reconsider whether the Pentagon should have a role in education. Government money once again seemed 'dirty'.

> *"Even money from private foundations comes with baggage."*

The stage was set for yet another funding model. Not long after the campus protests had faded, molecular biologists began to worry about the potential dangers of combining DNA sequences that don't occur together in nature. Cambridge emerged as one of the first cities in the United States to forge its own rules and procedures to allow such research. In short order, the area near MIT's campus became known as 'gene town', an incubator for private biotechnology companies, many of which enjoyed close ties to MIT faculty members and students.

The revolving door that has since existed between MIT life scientists, their students, corporate boards of directors and venture capitalists has surely been a great boon for research. But who benefits? Many critics fear that modern non-disclosure agreements are just as stifling as the corporate censorship rules of the Tech Plan or the defence department's classification codes. Other concerns loom as well. How much does MIT benefit when faculty members split their time between campus responsibilities and spin-off companies? Is private investment any more reliable or morally pure than public investment from the federal government?

Since the 1980s, MIT has followed a hybrid funding scheme, with clear roots in its earlier experiments. Now, about half of its annual budget comes from non-military branches of the federal government (the largest share from the Department of Health and Human Services); one-quarter from private industries and foundations; one-sixth from the military; and the remaining few per cent from state, local and foreign governments.

Even money from private foundations comes with baggage. The latest sparkling new building on MIT's campus — the David H. Koch Institute for Integrative Cancer Research, which officially opened last month — exemplifies the tensions. Billionaire businessman David Koch, an MIT alumnus and cancer survivor, generously funded the new building; he has donated comparable amounts to refurbish medical centres, museums and theatres across the country. Alongside his philanthropic giving, he has also funded conservative political groups associated with the 'Tea Party', although he denies any direct connection with that movement. Deserved or not, his name has become polarizing in today's political climate. At the building's grand opening, Koch declared that cancer is "absolutely non-partisan". True enough. But patronage, unlike the disease, is all about political choices and intellectual trade-offs.

Since MIT's founding, government sources and industrial sponsors have traded places several times, each held up alternately as saviour or poisoned fruit. As the institute's history has shown, no one model holds a monopoly on virtue. All patronage involves a delicate balance between opening up new opportunities and mortgaging intellectual autonomy. Rather than focus on the source of cash — public or private — we must remember to scrutinize the inevitable strings attached. ■

David Kaiser *is a physicist and historian at the Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. His latest book,* How the Hippies Saved Physics: Science, Counterculture, and the Quantum Revival *(W. W. Norton), will be published in June.*
e-mail: dikaiser@mit.edu

© 2011 Macmillan Publishers Limited. All rights reserved

Fix the antibiotics pipeline

As resistance mushrooms, governments must make development of new antibiotics financially viable for industry, say **Matthew A. Cooper** and **David Shlaes**.

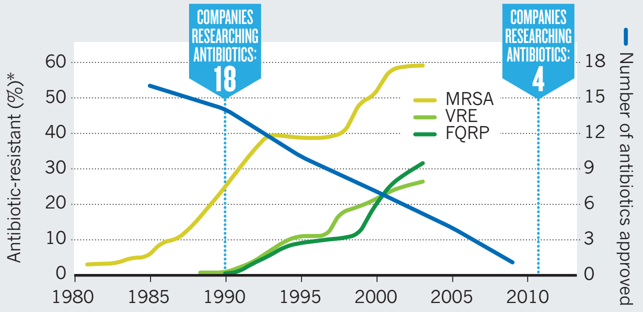
The framework for antibiotic discovery, development and approval is broken — only four new classes of antibiotics have been launched in the past 40 years. The World Health Organization forecasts a disaster due to the rapid, unchecked increase in antimicrobial resistance and has just announced a policy to combat its spread. But antibiotic resistance cannot be eliminated by stewardship alone. There needs to be a sustained effort from government and industry to develop new drugs quickly.

Phase III clinical trials — those on large groups of patients — of an antibiotic in a single disease indication cost about US$70 million. Funding for biotech companies — venture capital or government grants — cannot cover this for a drug that will be sold mainly in short courses, and to which resistance may emerge. Following the global recession, successful stock-market flotations of biotechnology firms remain rare. Many large pharmaceutical companies have abandoned research and development (R&D) on antibiotics, leaving few parties able to register and market new compounds.

Solutions have been debated over the past decade, but no concrete action has been taken. Before the end of 2011, the US government and European Union (EU) need to legislate a solution. Otherwise the hundreds of thousands of people dying each year from drug-resistant infections are likely to become millions.

A PERFECT STORM

As bacterial infections grow more resistant to antibiotics, companies are pulling out of antibiotics research and fewer new antibiotics are being approved.



*Proportion of clinical isolates that are resistant to antibiotic. MRSA, methicillin-resistant *Staphylococcus aureus*. VRE, vancomycin-resistant *Enterococcus*. FQRP, fluoroquinolone-resistant *Pseudomonas aeruginosa*.

SOURCE: CDC/IDSA

CARROTS AND STICKS

One solution was suggested in 2009 by the London School of Economics (LSE): a 'push–pull' incentive for investment in potential drugs that meet stringent criteria for medical need and probability of successful registration[1]. The 'push' would involve governments funding the otherwise prohibitively expensive phase III trials for at least one indication.

Push incentives lower the cost of market entry. They attract smaller enterprises with limited funds[2], but such firms may not have sufficient expertise to adequately manage phase III trials and get a drug to market. So a 'pull' is also required to engage larger companies with the necessary expertise and marketing reach.

One such pull suggested in the LSE report is guaranteed government purchase of a defined supply of the antibiotic for national stockpile — as happened for pandemic influenza and anthrax. Another pull is proposed by US congressman Henry Waxman (Democrat, California) in his bill Generating Antibiotic Incentives Now, which is with the House Committee on Energy and Commerce, pending US budget approval. This would give certain antibiotics five extra years of patent protection from generic competition. The bill would also enforce an expedited review of crucial new antibiotics by the US Food and Drug Administration (FDA) and encourages the FDA to designate life-saving antibiotics as a special regulatory class for priority review. Pull-only incentives promise financial rewards after a drug has been developed. Here the developer bears all the risk, because there is normally a decade or more between the decision to engage in R&D and commercial returns. The LSE push–pull incentive, plus Waxman's five-year patent extension, seems to us a clear front-runner. The promise of immediate (stockpile) and sustained (patent lifetime) revenues, plus subsidized phase-III development, would make it easier for small companies to go to the public market. It would also encourage the formation of new biotechnology ventures, providing a healthier climate for fundamental academic research.

Governments would get a significant return on investment. They would make savings, for example, in reductions in the estimated 2 million patients in the EU who catch hospital-acquired infections every year (of whom 175,000 die). Antibiotic resistance has been estimated to cost US hospitals more than $20 billion annually and add one to two weeks in hospital per patient. Subsidizing drug companies may be unpopular in many quarters, but it is necessary to bridge the gap between the high value of new antibiotics to society and the low returns they provide to drug companies.

LEADERSHIP NEEDED

Since 2006, the FDA has demanded more costly, larger-cohort studies to prove the non-inferiority of a candidate drug over an existing antibiotic. Without change at the FDA, antibiotic developers, especially smaller companies, may simply ignore the agency. If the European Medicines Agency (EMEA) continues to allow swift, affordable trial designs for key antibiotic indications, a company could use approval in Europe to drive approvals in growing markets such as India, China and Brazil. For example, Johnson & Johnson's doripenem can be used to treat nosocomial pneumonia in almost all countries except the United States.

The Trans Atlantic Task Force on Antimicrobial Resistance prepared a draft proposal[3] for the EU and United States defining areas of future cooperation and policy alignment between industry and governmental agencies. This proposal and the push–pull model above require urgent translation into policy. Antibiotic resistance is a global health crisis. It requires global action before one of the most valuable scientific discoveries of the twentieth century is lost in the twenty-first century. ■

Matthew A. Cooper *is a professor at the Institute for Molecular Bioscience, University of Queensland, Brisbane St Lucia, QLD 4072, Australia.* **David Shlaes** *is a former vice-president at Wyeth Pharmaceuticals, and author of* Antibiotics: The Perfect Storm *(Springer).*
e-mails: m.cooper@imb.uq.edu.au; shlaes.david@earthlink.net

1. Mossialos, E. *et al. Policies and Incentives for Promoting Innovation in Antibiotic Research* (London School of Economics and Political Science, 2009).
2. Morel, C. M. & Mossialos, E. *Br. Med. J.* **340,** c2115 (2010).
3. Report on the European Commission's Public Online Consultation Stakeholder Consultation on the Transatlantic Task Force on antimicrobial resistance available at http://go.nature.com/hqetfw

D. Shlaes declares competing financial interests: details accompany the full-text HTML version of the paper at go.nature.com/bogtg3.

© 2011 Macmillan Publishers Limited. All rights reserved

D. ECKSTEIN/ZUMA PRESS/NEWSCOM

Author Joshua Foer winning the USA National Memory Championship in 2006.

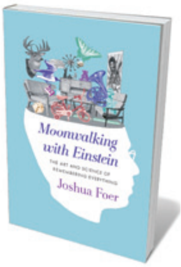
PSYCHOLOGY

The art of remembering

Larry R. Squire enjoys an engaging account of how memory works and how to win in memory competitions.

In ancient times, before printing and the ready availability of paper for taking notes, a trained memory was of utmost importance. In his writings on oratory, the Roman statesman and philosopher Cicero recounted how the art of memory was invented 400 years earlier in Greece and became a corpus of techniques that were widely known and widely practiced. The method of loci, as it is commonly known, involves constructing visual images for items to be remembered, and then placing them at spatially discrete locations in a familiar building or along a known path. The items are then recovered from memory by mentally retracing one's steps. The only existing classical treatise on the subject (*Ad Herrenium*, written about 86–82 BC) and the most authoritative modern treatment (Frances Yates' *The Art of Memory*, Routledge and Kegan Paul; 1966) describe the technique and trace its history to the seventeenth century.

The best-known modern example of this method at work comes from Alexander Luria's *The Mind of a Mnemonist* (Basic Books, 1968). Luria, a Russian neuropsychologist, recorded the remarkable ability of a man identified as S, who from his early life had an essentially unlimited memory capacity. He could listen to long lists of words or numbers and later give back the whole list, as long as he had a few seconds to visualize each item at the time of learning. These images, which were aroused involuntarily in response to sensory impressions, were then placed along a familiar street. When, infrequently, he omitted an item, the difficulty was in perception, not in memory. For example, he describes why he once omitted the word egg: "I had put it up against a white wall and it blended in … How could I possibly spot a white egg up against a white wall?" S's abilities came with serious drawbacks. His memory was so cluttered with detail and so overwhelmed with separate images that he could not notice the regularities among related experiences. Metaphor and poetry were often beyond him. For S, his ability was abnormal and a tragic burden.



Moonwalking with Einstein: The Art and Science of Remembering Everything
JOSHUA FOER
Penguin/Allen Lane: 2011. 320 pp. $26.95/£14.99

In his charming book *Moonwalking with Einstein*, journalist Joshua Foer describes how the same extraordinary feats of memory are possible through training, even for people with average memories. These are the mental athletes who compete in memory championships and whose achievements are scarcely credible: memorizing lists of binary digits (record: 4,140 digits in 30 minutes); the order of a deck of playing cards (record: 32 seconds); as many decks of playing cards as possible in one hour (record: 27 decks).

The book is interwoven with informed exposition about the psychological science of memory (including discussion of one of this reviewer's study patients). What happens when memory fails, as in amnesia after brain injury? What is the nature of acquired expertise, such as typing, ice skating or the reading of mammograms? Studies show that recently trained radiologists are more accurate at screening mammograms than those who were trained some time ago. Because the accuracy of a diagnosis becomes known only later, practitioners may forget the case details and so cannot learn effectively from their mistakes. As a result, their skills can backslide.

Experts engage in directed, highly focused practice. In the words of the great American-football coach, Vince Lombardi: "Practice does not make perfect. Perfect practice makes perfect." The best ice skaters spend more time practising routines they have difficulty with, whereas lesser skaters work more on routines they have mastered. Typists can improve their speed by deliberately practising at a rate above whatever plateau they have reached and then analysing their errors.

The most entertaining parts of the book are the detailed accounts of how the mental athletes prepare for competition and develop expertise, including the author himself, who in a one-year adventure entered and won the US Memory Championship. The technique begins by selecting a familiar space such as a large building (historically called a memory palace), which the competitor populates with the images of what is to be remembered. Experts use dozens of these, each with a unique route that can be followed for depositing and retrieving images. The images themselves are crucial — they should be highly detailed, bizarre, even lurid. Competitors need to deliberate on each image to know its colour and shape, its smell and texture, and their feelings about it. When the image is a person, the introduction of humour, action and sex is encouraged.

The main competitive method for remembering the order of a deck of playing cards involves pre-memorizing a

➲ NATURE.COM
For a review of Oliver Sacks's *The Mind's Eye*, see:
go.nature.com/cshza4

© 2011 Macmillan Publishers Limited. All rights reserved

subject–verb–predicate image for each card. For example, the king of diamonds: Dad riding a tricycle or, using the book's title, Dad moonwalking with Einstein. In tournaments, cards are turned over three at a time. A new image is then constructed to represent all three cards (images ABC, DEF and GHI are converted into the single image AEI). This new image retains a trace of each card's identity, which can be placed in a room of the memory palace for later retrieval. To support memory of a 52-card deck, the expert needs 17 images plus one additional image.

Because a collection of images typically includes a number of titillating acts, difficulties can arise when combining them results in images of family members engaged in socially unacceptable practices. As he prepared for competition, Foer worried that he was being distracted by the indecent acts his mother had to commit "in the service of my remembering the eight of hearts". His coach knew the problem: "I had to excise my mother from the deck. I recommend you do the same."

After gaining his trophy, Foer retired from competition and now rarely uses the techniques. They are effortful and hardly vital in an age of external memory in which remembering information may be less important than knowing how to find it. Also, ordinary memory works at cross-purposes with memory-training techniques. We are best at generalizing, abstracting and assembling general knowledge, not at retaining a literal record of events: we forget the particulars and thereby can retain the main points. Studies show that people will remember the meaning of sentences but forget whether the sentences were presented in the active or passive voice. Freud wrote: "Normal forgetting takes place by way of condensation. In this way it becomes the basis for the formation of concepts."

Influenced by these realities of memory, current pedagogy has minimized rote memorization and drills, emphasizing instead problem solving and independent thinking. Yet, if it is true (as stated in the book) that two-thirds of US teenagers cannot locate the Civil War within 50 years, or that 20% cannot identify the adversaries in the Second World War, perhaps there is a place in education for the skill of memorizing. Foer tells of an inner-city history teacher, an enthusiast of memory training, who introduced the techniques rigorously, comprehensively and with considerable success. As Foer writes, "even if facts don't themselves lead to understanding, you can't have understanding without facts." ■

Larry R. Squire *is professor of psychiatry, neurosciences and psychology at the University of California, San Diego, and research career scientist at the Veterans Affairs Healthcare System, San Diego, California 92161, USA.*
e-mail: lsquire@ucsd.edu

U. SINAI/GETTY

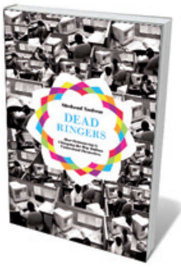
The time difference with the West takes its toll on a web tutor at an Indian-based company.

GLOBALIZATION

Behind India's technological boom

The rise of outsourcing by Western companies stifles local innovation, learns **Andrew Robinson**.

It is now an everyday experience to phone a large US- or UK-based company with a technical, financial or administrative enquiry and end up talking to someone in Bangalore or Mumbai. India's ready supply of well educated, English-speaking and relatively cheap workers has made the country a top destination for many Western companies, from banks and airlines to big pharma and information technology (IT) firms. Yet the outsourcing of labour has had unforeseen local impacts on science and innovation — and on the technologically gifted young people of India.



Dead Ringers: How Outsourcing is Changing the Way Indians Understand Themselves
SHEHZAD NADEEM
Princeton University Press: 2011. 288 pp. £24.95

Dead Ringers, by US sociologist Shehzad Nadeem, is the first academic field study to explore what turns out to be an occupational dead end for hundreds of thousands of Indians working for Western corporations. As well as the multitude who man telephones, this vast group includes an army of software programmers, accounting specialists and interpreters of medical scans. Nadeem interviewed more than 125 workers, managers, employers and trade unionists in India and the United States, mainly in 2005–06. He offers concrete and important insight into the world of outsourcing, but in highlighting the downsides, he downplays the undeniable successes and the homegrown roots of India's research and development sector.

India's IT boom, which started in the mid-1990s after the liberalization of the Indian economy in 1991, has generated headlines and hyperbole in both business and politics. As Nadeem readily accepts, outsourcing has provided many young Indians with comparatively well-paid opportunities and it has boosted India's reputation internationally. In 2004, the boom even contributed to the electoral slogan of the ruling Bharatiya Janata Party (BJP), "India shining".

But the BJP's controversial phrase turned out to be ill chosen. The party lost

© 2011 Macmillan Publishers Limited. All rights reserved

the general election, and the IT boom began to lose its shine, especially after admissions of false accounting in 2009 led to the collapse of Satyam Computer Services, which in 1999 was one of the first India-based IT companies to be listed on the NASDAQ stock market. It is now widely recognized that the 1990s dream of Indian development led by outsourcing was, in Nadeem's words, "wildly oversold".

The offices of India's glamorous IT companies may look "like twinkling towers of innovation", says the author. But he contends that "like plastic fruit, they are imitations". Nadeem backs up this view with skilfully told stories from his Indian sources (some named, most anonymous). But when he tries to uncover the reasons why India's IT industry has generally failed to innovate at home, despite prominent individual successes among Indians in California's Silicon Valley and Western academia, his conflation of outsourcing with research and development blurs his analysis and conclusions. The book is also heavy on academic jargon.

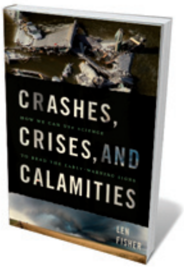
CREATIVITY CURTAILED

What does he think is stalling homegrown innovation? A lack of emphasis on individualism in Indian family life and a widespread deference to authority may be part of the reason, says Nadeem. But much more important, he says, is the global economic system. Outsourcing managers, both in India and the United States, are "locked in a contradiction", he notes. They want their workers to mature into professionals who show initiative and take responsibility for projects. But simultaneously, they want to migrate easily replicable, standardized tasks rather than whole projects to India. Farming out tasks generates a reliable stream of revenue while ensuring that control of the process remains based in the United States and Europe. The core work requiring creativity therefore stays in the West.

Nor have these elite industries been effective in alleviating India's massive poverty, Nadeem argues, despite generating impressive economic growth. Between 1994 and 2008, India's export revenues in IT and IT-enabled services (ITES) grew from less than US$0.5 billion to $40.4 billion, and are predicted to reach $71 billion in 2011. Between 2004 and 2008, the number of workers employed in the sector — most of whom are male and in their twenties — more than doubled, to an estimated two million. Their average entry-level salary in IT in 2006 was $5,715 a year (compared with $46,194 in the United States). By 2007, India's IT and ITES industry accounted for 5.2% of the country's gross domestic product. Yet, the IT and ITES

➔ NATURE.COM
A review of a book on globalization and education:
go.nature.com/zov6dv

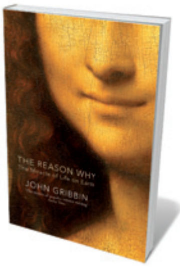
Books in brief



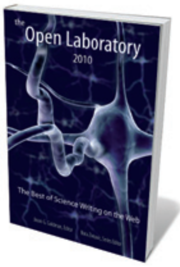
Crashes, Crises, and Calamities: How We Can Use Science to Read the Early-Warning Signs
Len Fisher Basic Books *256 pp. £13.99 (2011)*
From earthquakes to the collapse of civilizations and economies, why do systems suddenly break down? Physicist and writer Len Fisher gives an accessible explanation of the mathematics of catastrophes in his latest book. Drawing on physics, ecology and biology, he highlights four tools that scientists and engineers use to forecast rapid failure: stability, catastrophe, complexity and game theory. By applying these concepts, he explains, we can predict and manage impending crises.

Science-Mart: Privatizing American Science
Philip Mirowski Harvard University Press *464 pp. $39.95 (2011)*
Since the 1980s, commercial companies have become the largest funders of scientific research in the United States. Economist and historian Philip Mirowski analyses in detail the impact of this shift away from public funding. Owing to the rise of patents and intellectual property, knowledge and discovery are now perceived as a commodity; the fruits of scientific investigations are no longer considered a public good but are seen as products with monetary value. But, the author argues, American science should be more than just a cash cow.

Nightwork: A History of Hacks and Pranks at MIT
T. F. Peterson MIT Press *232 pp. $22.95 (2011)*
Every university has its canon of creative pranks, which usually involve a degree of technological know-how. The finest practical jokes played at the Massachusetts Institute of Technology in Cambridge are documented in this illustrated volume (first published in 2003), which has been updated for the institution's 150th anniversary. Recent 'hacks', as they are known, include the cross-country theft of a cannon from the California Institute of Technology in 2006, and the hoisting of a solar-powered subway car and a fire engine onto the roof of the university's Great Dome.



The Reason Why: The Miracle of Life on Earth
John Gribbin Allen Lane *240 pp. £20 (2011)*
A series of one-off cosmic events and flukes of physics represent the lucky breaks that made our planet the oasis it is today, argues best-selling writer John Gribbin in his latest book, which examines the origin of life on Earth. From the giant collisions of early Solar System bodies that forged our planet to the geochemical reactions that make it habitable, our world is special. Even though planets are common in the Milky Way, Gribbin argues, intelligent life capable of building technological civilizations will turn out to be rare. So the future of humankind is of universal significance.



The Open Laboratory 2010: The Best of Science Writing on the Web
Edited by Jason Goldman & Bora Zivkovic Lulu *284 pp. £11.91 (2011)*
Last year saw the eruption of Iceland's Eyjafjallajökull volcano, the deep-water oil spill in the Gulf of Mexico and the announcement of arsenic-based life. It was also, according to Jason Goldman, editor of this collection of 2010 blog posts, the year when blogging went mainstream. Thanks to the extra boost of Twitter, online diarists were sought out to comment on all the big science stories. Their insights are shared in this annual selection of the best of the blogosphere.

© 2011 Macmillan Publishers Limited. All rights reserved

workforce — which is frequently located in deregulated special economic zones that are cut off from their surroundings — constitutes less than 0.5% of India's total workforce of some 450 million, 92% of whom survive as labourers, farmers and street vendors. According to a 2007 Indian government report, 77% of Indians live on less than 50 cents a day. In 2010, only 366 million Indians had access to modern sanitation (for comparison, India has 564 million mobile phones).

Nadeem gives vignettes of life in four unnamed outsourcing companies in Bangalore, Mumbai, Chennai and New Delhi, where long hours, graveyard shifts and stressful monitoring regimes without doubt damage the workers' health. These snapshots, combined with the start-ups that have failed to live up to their promise, and widespread corruption in Indian business and politics, give the 'dead' in *Dead Ringers* — which initially refers to call-centre workers' dubious mimicry of Western accents — a more ominous significance as the book progresses.

Nadeem concludes that outsourcing is a new form of colonialism, with an insidious appeal for young Indians in thrall to American mass consumerism. Although that is essentially true, his simple explanation skirts the internal impetus that has been given to Indian technological innovation. After India became independent in 1947, its first prime minister, Jawaharlal Nehru, did much to establish the country's scientific higher-education system, including the Indian Institutes of Technology, and to build up its technological sector, on which the success of the IT industry rests. Nadeem neglects this crucial background and seems to endorse an unnamed Indian executive's dismissive comment: "The only thing that Nehru gave us was education. That allowed people to be in a good position when the knowledge boom came."

"The brave new IT world documented in Nadeem's interviews disturbs more than it shines."

There is more variety and originality in Indian IT than *Dead Ringers* implies. Nonetheless, for all the wealth and political prestige that outsourcing has brought to India, one cannot help agreeing with the author that the brave new IT world documented in his interviews disturbs more than it shines. ■

Andrew Robinson *is a London-based author. He is the biographer of Indian film-maker Satyajit Ray and Nobel laureate and writer Rabindranath Tagore.*
e-mail: andrew.robinson33@virgin.net

ASTRONOMY

Finding other worlds

A survey of exoplanetary research shows how the field has come in from the cold, finds **Chris Tinney**.

When I began my career as a graduate student in astronomy in the late 1980s, it was clear which fields were considered hot, which were not, and which were outré. Cosmology and infrared astronomy were hot. Galactic dynamics and most stellar astronomy were staid. And the search for planets and brown dwarfs — the class of objects intermediate in mass between planets and stars — was definitely outré. How things have changed.

In *Strange New Worlds*, astronomer and one-time journalist Ray Jayawardhana surveys how 15 years of exoplanet discovery has changed astrophysics. From a small base, in terms of personnel and funding, exoplanetary science — the search for and study of planets orbiting other stars — has grown rapidly and now sits at the core of modern astrophysics. Its findings have overturned many established ideas.

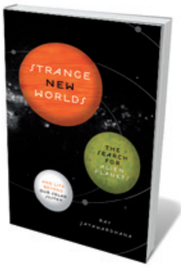
To show how far we have come, Jayawardhana relates a telling incident from the late 1980s: a distinguished astronomer strode out of the room when a pioneer of exoplanet searches, Gordon Walker, rose to speak about his work. As Walker remarks, it "seems hard to believe now".

Nonetheless, a few brave souls continued to work in the field. Advances in astronomical detectors, instrumentation and analysis techniques meant that in 1995, hundreds of years of observations finally bore fruit and the signatures of orbiting planets were discovered in the spectra of other stars. In the years since, hundreds of exoplanets have been found. The change in the landscape of astronomy and planetary science has been profound.

Entirely new fields have come into being or come into their own: exoplanetary science, and astrobiology, which explores the possibility of life elsewhere in the Universe. New research groups have popped up around the globe, backed by governments through major scientific strategies. The first sentences of the executive summary of the 2010 US National Research Council's astronomy decadal survey highlight this shift: "Our view of the universe has changed dramatically. Hundreds of planets of startling diversity have been discovered orbiting distant suns." Astronomy's old hierarchies have been overturned.

The title of Jayawardhana's book reflects the major scientific finding of all this activity: exoplanets are much stranger than we expected. Very few of the planetary systems found around other stars resemble the architecture of our own Solar System. Most exoplanet orbits are highly elliptical and near-circular orbits are rare, occurring only when gravitational tidal effects make them so. Gas- and ice-giant planets are located in places where they could not originally have formed, indicating that they have moved great distances since formation. This migration seems to be the dominant driver of the exoplanet architectures we observe.

➲ NATURE.COM
For more on the search for exoplanets:
go.nature.com/hjdqvb



Strange New Worlds: The Search for Alien Planets and Life Beyond Our Solar System
RAY JAYAWARDHANA
Princeton University Press: 2011. 288 pp. £16.95, $24.95

Some exoplanets are much denser than expected; others are much less dense. Some systems host many ice-giant planets in tight orbits, whereas our Solar System has only one tiny terrestrial planet (Mercury) so close. Others host no giants at all. Evidence is beginning to emerge that Earth-like, or terrestrial, planets might not be the norm. The Copernican principle — that Earth is not special or unusual — may not hold after all.

Jayawardhana's presentation of the research is remarkably even-handed. This is a fast-moving field in which groups have often clashed. Nonetheless, he provides a survey of the subject without giving the protagonists anything to complain about. His lucid and effortless prose makes for an engaging read.

Strange New Worlds anticipates the major results that can be expected in exoplanetary science in the coming decades: the imaging of exoplanets orbiting nearby stars; finding the first habitable Earth-like planets; the detection of biomarkers that suggest the existence of life outside the Solar System. These and much more will continue to make this field not just fashionable, but very exciting. No one is walking out of the room any more. ■

Chris Tinney *is professor of exoplanetary science at the University of New South Wales, Sydney, Australia.*
e-mail: c.tinney@unsw.edu.au

© 2011 Macmillan Publishers Limited. All rights reserved

E. M. PEARSON/ERICMICHAELPHOTOGRAPHY.COM

At a Story Collider show, Naomi Azar relates how her brain disorder stops her from recognizing faces.

COMMUNICATION

Show and tell

A New York storytelling project reveals the personal side of scientific life, finds **David Carmel**.

The man behind the microphone is clearly nervous. He is a neuroscientist, but this is not a conference or lecture. He is telling a deeply personal story at The Story Collider, a monthly New York event that has been gaining popularity for the past year, and which attracts a diverse audience to hear stories about science.

On 13 April, The Story Collider joins forces with the Zora Art Space in Brooklyn to produce a special event and exhibition of art, science and storytelling titled *The Cambrian Explosion*. On the opening night, four contemporary artists and two curators will tell stories about how science has influenced their work and touched their lives. Their work — paintings, photographs and sculptures — will then be exhibited until 28 April. "This will be an energetic show," promises co-curator Eric LoPresti, "dissecting neuroscience, biology, psychology and evolution through smart, visually stunning artwork."

Artists Karen Margolis and Elizabeth Demaray have undergraduate degrees in cognitive science. Margolis will show maps and 'architectural renderings' that portray emotional states and mental operations, which she created using a soldering iron to burn intricate patterns resembling neural networks into layers of pigmented paper. Sculptor Demaray will install her Lunch Box Project, which invites audience members to contribute food scraps to a population of 3,000 earthworms housed in a large, clear-plastic vitrine, creating a time-based 'painting' in which leftovers produce nutrient-rich soil.

The Cambrian Explosion
Zora Art Space, New York. 13–28 April 2011.
http://storycollider.org

Joining Margolis and Demaray are photographer Lori Nix, who will present pictures of elaborate constructed scenes — an extension of her solo show in Chicago in February, which featured a future city emptied of its human inhabitants — and painter Kysa Johnson. Johnson will exhibit monochromatic chalk drawings of luscious American landscapes, which on close inspection reveal themselves to be composed not of rivers and foliage, but of depictions of the molecular structures of the environmental pollutants ethane, benzene and acrolein.

Conversations between science and the arts are notoriously difficult. They are often hampered by the incompatibility of academic discourse on the two sides. "The idea behind this show," says co-curator and physics PhD Ben Lillie, "is that letting the artists tell their stories of how science has affected them can be a grounding point — a more directly human way of starting the conversation."

This is The Story Collider's first foray into bridging art and science. In the future, Lillie plans to do the opposite show, with scientists telling stories about how art has influenced their research. Lillie co-founded The Story Collider with Brian Wecht, a postdoctoral physicist at the University of Michigan in Ann Arbor. They were looking for a way to let people share their personal experiences of science.

There are many popular-science events in New York — Café Scientifique, The Secret Science Club and Nerd Nite, to name a few. These focus on imparting knowledge. Storytelling shows are another hugely successful cultural trend in the city, with events such as The Moth, The Liar Show and Told. The Story Collider combines both trends. "There are no lectures," says Lillie. "I want our audience to get a sense that science is a part of people's lives."

In The Story Collider's regular monthly shows, which take place at the Pacific Standard venue in Brooklyn, six performers tell stories about the part science has played in their lives, without notes or props. Each evening has a broad theme; past events have been titled 'Parallel Universes', 'Standard Deviation' and 'My Science Project'. They yield an eclectic mix of stories: Columbia University cosmologist Eugene Lim recounted his experiences on an aid trip to Haiti after last year's earthquake; comedian Dave Ritz told the audience how it felt to be accused of stealing, in what turned out to be a psychology experiment; and performer Seth Lind described living with a pacemaker.

"What I like most is that you can hear how science has affected absolutely everyone," says Lillie. "Scientists are often asked to talk about their work, but rarely about themselves."

This is not an entirely impartial review. That neuroscientist storyteller at the start of this report? That was me. February's theme was cognitive dissonance, and I told the audience how my teenage encounter with Oliver Sacks's books about fascinating neurological syndromes influenced my career choice, and how, years later, I was torn between scientific excitement and intense worry when my own father developed bizarre symptoms following a stroke.

Telling my story was nerve-racking at first, but as the minutes passed I could feel the crowd's attentiveness and sympathy, and I became more confident. Overall, the experience was incredibly cathartic, and having people come up to me afterwards and tell me how interesting and moving they found my story was as gratifying as any compliment I have ever received about my research. Storytelling is definitely a unique way to communicate the human side of scientific life. ■

David Carmel *is at the Department of Psychology and Center for Neural Science, New York University, New York 10003, USA.*
e-mail: davecarmel@nyu.edu

© 2011 Macmillan Publishers Limited. All rights reserved

CORRESPONDENCE

China's green policy has some way to go

The ambitious environmental and energy targets set out in China's latest five-year plan (*Nature* **471,** 149; 2011) should be considered in relation to the economic realities of environmental decline and of China's governance.

Lowering growth targets for gross domestic product (GDP) and focusing on environmental and energy issues should not be viewed as a complete shift away from the 'economy first' paradigm that has driven China's national agenda for the past 30 years. The new policies are aimed at a more socially inclusive view of economic development, in line with recent criticisms of GDP as a measure of social welfare (see, for example, go.nature.com/to4ppq).

Despite China's remarkable growth rate, a report released by its government in 2006 revealed that roughly 3% of the country's annual GDP had been offset by economic loss through environmental degradation — a figure that some think is too conservative (*Nature* **448,** 518–519; 2007). China's present and future environmental policies should continue to acknowledge the high economic cost of environmental problems in sustainable development.

Tong Wu *Northern Arizona University, Arizona, USA.*
tong.wu@nau.edu

Financial model failed in real world

David Lindley cites the Black–Scholes model as a means of calculating and predicting stock-market variations (*Nature* **471,** 255–256; 2011). But the model has its pitfalls.

For example, the US hedge-fund firm Long-Term Capital Management used this approach to direct its fund. The fund crashed in 1998 because the predictions diverged sharply from reality. The US government had to bail the firm out at a cost of about US$4 billion.

Robin O. Motz *New York Academy of Sciences, USA.*
robinmotz@gmail.com

An insight into Maxwell's mind?

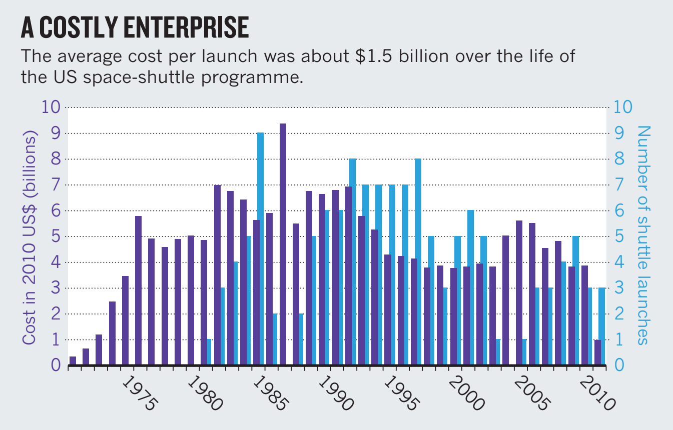
Among the comments on the 150th anniversary of James Clerk Maxwell's groundbreaking paper *On Physical Lines of Force* (*Nature* **471,** 289–291; 2011), nothing was said about what drove the thinking of this great physicist.

Personal perspectives can provide insight into the dynamics of scientists' behaviour. As the Victorian age matured, science leaders became increasingly materialistic. At a meeting of the British Association in 1874, president John Tyndall took the opportunity to advance his world view of materialism. Maxwell was in the audience and crafted a poem to express his disquiet, the first verse of which runs:

> In the very beginnings of science, the parsons, who managed things then,
> Being handy with hammer and chisel, made gods in the likeness of men;
> Till Commerce arose, and at length some men of exceptional power
> Supplanted both demons and gods by the atoms, which last to this hour.
> Yet they did not abolish the gods, but they sent them well out of the way,
> With the rarest of nectar to drink, and blue fields of nothing to sway.
> From nothing comes nothing, they told us, nought happens by chance, but
> by fate;
> There is nothing but atoms and void, all else is mere whims out of date!

Such thorny issues may well have influenced Maxwell's perspective on science.

David J. Tyler *Manchester Metropolitan University, UK.*
d.tyler@mmu.ac.uk



A COSTLY ENTERPRISE

The average cost per launch was about $1.5 billion over the life of the US space-shuttle programme.

Shuttle programme lifetime cost

With the final two flights of the NASA space-shuttle programme scheduled for later this year, we can now evaluate the lifetime cost of the programme.

Some 20 years ago, we found the programme to be slightly over budget and severely short in capability (R. A. Pielke and R. Byerly in *Space Policy Alternatives* Ch. 14, 223–245; 1992). We used 8 years of cost and schedule experience to predict performance for the subsequent 20 years of the shuttle programme.

The US Congress and NASA spent more than US$192 billion (in 2010 dollars) on the shuttle from 1971 to 2010 (see 'A costly enterprise'). The agency launched 131 flights; two ended in tragedy with the loss of *Challenger* in 1986 and *Columbia* in 2003. During the operational years from 1982 to 2010, the average cost per launch was about $1.2 billion. Over the life of the programme, this increases to about $1.5 billion per launch (R. A. Pielke *Space Policy* **10,** 78–80; 1994).

For the period 1991–2010, we originally projected an average cost per flight of about $800 million. The actual cost was about $1 billion. We overestimated both the flight rate during this time (8 predicted flights versus 4.7 actual) and the annual costs (about $6.2 billion predicted versus $4.7 billion actual).

The actual cost for each flight of the programme falls squarely in the middle of the envelope we constructed, with projected uncertainties. Thus, our 1992 projection indicates that the performance of large-scale technologies might be predictable if projections are based on past experience.

The shuttle is the costliest US spaceflight programme ever undertaken. As it comes to an end, we should celebrate its successes, and draw lessons to inform future human spaceflight ventures.

Roger Pielke Jr, Radford Byerly *University of Colorado, Boulder, USA. pielke@colorado.edu*

SEE COMMENT P.27

CONTRIBUTIONS

Correspondence may be sent to **correspondence@nature.com**; guidelines are at **go.nature.com/cmchno**.

© 2011 Macmillan Publishers Limited. All rights reserved

CARDIOVASCULAR DISEASE

The diet–microbe morbid union

A common dietary component that some people even take as a supplement is converted by the gut microbiota to harmful metabolites linked to heart disease. This finding has cautionary implications. SEE ARTICLE P.57

KIMBERLY RAK & DANIEL J. RADER

Everyone knows that a 'bad diet' can lead to heart disease. But which dietary components are the most harmful? Some lay the blame on saturated fatty acids, others point a finger at excess carbohydrates, which also lead to obesity and insulin resistance. On page 57 of this issue, Wang *et al.*[1] outline a remarkable chain of events that links diet, intestinal bacteria and liver metabolism to the generation of a chemical that promotes the build-up of arterial plaque and cardiovascular disease.

Intestinal bacteria currently hold centre stage for their role in maintaining digestive health[2]. Although the main focus has been on detailed molecular characterization of the gut microbiome[3], there is increasing interest in the impact of these seemingly innocuous gut microorganisms on metabolic disease in humans. Indeed, recent evidence[4,5] has implicated gut microbiota in insulin resistance and non-alcoholic fatty-liver disease.

A burgeoning area of research is metabolomics — an unbiased approach to identifying and measuring the small-molecule metabolites in a system — and determining the relationship of the metabolome to disease. Nonetheless, the scope of the blood metabolome that arises from the gut microbiome has not been fully defined; this knowledge could lead to insights connecting diet, the gut microbiota and disease.

Wang *et al.*[1] tell a compelling story — which starts with a metabolomics approach — of their search for circulating small molecules associated with coronary heart disease. They screened blood from patients who had experienced a heart attack or stroke and compared the results with those from blood of people who had not.

The authors found major differences in choline, betaine and trimethylamine *N*-oxide (TMAO) — three metabolites of the ubiquitous dietary lipid phosphatidylcholine (also called lecithin). Choline is an essential nutrient[6], and lack of dietary choline can lead to non-alcoholic fatty-liver disease and muscle damage. This knowledge has sparked the use of choline as a dietary supplement to prevent liver damage and to increase muscle performance[7]. Furthermore, lecithin is marketed as a dietary supplement to reduce the risk of heart disease, despite the absence of data to support this claim.

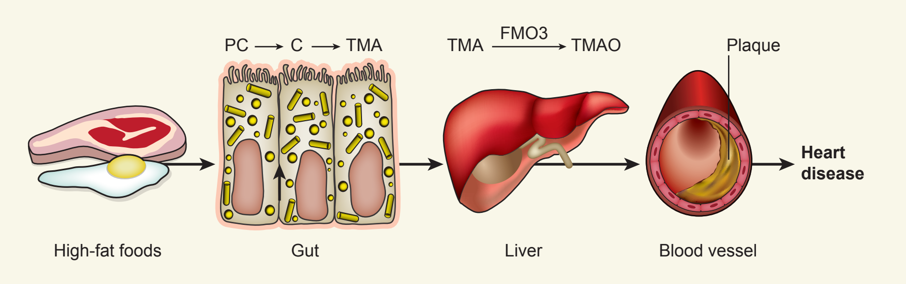


Figure 1 | From diet to disease. High-fat foods are rich in the lipid phosphatidylcholine (PC) and its metabolite choline (C). Intestinal bacteria convert C to TMA. In the liver, the enzyme FMO3 processes TMA to TMAO — a metabolite that makes its way into the blood. Wang *et al.*[1] show that circulating TMAO may contribute to greater plaque development in the arteries, and so to heart disease.

After choline is released from phosphatidylcholine by phospholipase enzymes, gut microbiota metabolize much of it into trimethylamine (TMA) — a gas that smells like rotten fish. When TMA reaches the liver, oxidizing flavin monooxygenase enzymes convert it into TMAO (Fig. 1). Wang and colleagues gave mice isotopically labelled phosphatidylcholine orally and later found the label in TMAO in the animals' plasma, confirming the metabolic link between dietary intake of phosphatidylcholine and the production of TMAO. They also report that, in mice prone to atherosclerosis, increased dietary choline leads not only to increased plasma levels of TMAO, but also to greater plaque development in the animals' arteries.

To demonstrate the role of gut microbiota in this process, the researchers treated mice with broad-spectrum antibiotics, effectively abolishing the animals' intestinal flora. In this setting, phosphatidylcholine administration did not result in TMAO in the blood and, more strikingly, a high-choline diet did not increase the severity of atherosclerosis. A lingering question is whether increased TMAO production contributes to cardiovascular disease or is simply a marker of disease risk.

This paper[1] raises the possibility of several new approaches to prevent or treat atherosclerosis. The most obvious is to limit dietary choline intake. Although phosphatidylcholine is found in a wide range of foods, it tends to be particularly high in foods with greater fat content[6]. Indeed, people with trimethyluria, who cannot convert TMA to TMAO, are prescribed a low-fat, low-choline diet to reduce TMA production[8]. What's more, our diet comparisons show that a very low carbohydrate (Atkins) diet contains roughly 2.5 times more choline than a typical very low fat (Ornish) diet. It is thus tempting to speculate that a very low fat diet may reduce the risk of heart disease in part because of its low choline content (Fig. 1). These results also call into question the safety of using choline and lecithin as dietary supplements.

Another approach is to reduce the load of gut bacteria that generate TMA from dietary choline. Intriguingly, low-dose antibiotics have been used[8] to reduce TMA production in people with trimethyluria. It is of note that antibiotic trials in humans[9] set up to test the hypothesis that certain microorganisms, such as chlamydia, may directly infect the arterial wall have not shown cardiovascular benefit. If bacterial species responsible for metabolizing choline to TMA are identified, their selective elimination would be ideal because it would be therapeutically sufficient, and less disruptive to the intestinal microbiota than the broad-spectrum antibiotics Wang *et al.* used in mice.

A third approach is to use probiotics — live microorganisms that both inhibit and promote various species in the gut microbiome. In a mouse model carrying a 'humanized' microbiome[10], administration of a certain probiotic reduced TMAO production, whereas another probiotic increased it. Clinical studies

© 2011 Macmillan Publishers Limited. All rights reserved

of the effect of probiotics on plasma TMAO levels and on cardiovascular disease in humans would be of interest.

Because TMAO is produced in the liver by the action of the flavin monooxygenase FMO3, inhibition of this enzyme in the liver might be another strategy by which to reduce TMAO production and cut the risk of heart disease. Although complete absence of FMO3 — for instance, in the disease trimethylaminuria — is undesirable, its reduced activity might be beneficial. Whether variations in the gene encoding FMO3 that reduce its activity are associated with reduced plasma TMAO levels and, more importantly, with reduced incidence of cardiovascular disease, should be tested.

Although Wang and colleagues' work[1] suggests that excess dietary choline might lead to cardiovascular disease, choline is an essential nutrient for several cellular metabolic pathways. So any attempt to reduce the levels of choline or its metabolites for therapeutic purposes requires caution. Nonetheless, this study has added phosphatidylcholine and other sources of dietary choline — such as the widely used food supplements — to the list of dietary culprits with the potential to increase the risk of heart disease. What's more, it implicates the gut microbiome in promoting heart disease in the setting of a high-choline diet. The implications for prevention of cardiovascular disease are tangible, and the subsequent chapters in this story will make fascinating reading. ■

Kimberly Rak *and* **Daniel J. Rader** *are at the Institute for Translational Medicine and Therapeutics, and the Cardiovascular Institute, University of Pennsylvania Schools of Medicine and Veterinary Medicine, Philadelphia, Pennsylvania 19104-6160, USA.*
e-mail: rader@mail.med.upenn.edu

1. Wang, Z. *et al. Nature* **472,** 57–63 (2011).
2. Chow, J., Lee, S. M., Shen, Y., Khosravi, A. & Mazmanian, S. K. *Adv. Immunol.* **107,** 243–274 (2010).
3. Gill, S. R. *et al. Science* **312,** 1355–1359 (2006).
4. Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. *Science* **307,** 1915–1920 (2005).
5. Dumas, M.-E. *et al. Proc. Natl Acad. Sci. USA* **103,** 12511–12516 (2006).
6. Zeisel, S. H., Mar, M.-H., Howe, J. C. & Holden, J. M. *J. Nutr.* **133,** 1302–1307 (2003).
7. Block, G. *et al. Nutr. J.* **6,** 30 (2007).
8. Busby, M. G. *et al. J. Am. Diet. Assoc.* **104,** 1836–1845 (2004).
9. Andraws, R., Berger, J. S. & Brown, D. L. *J. Am. Med. Assoc.* **293,** 2641–2647 (2005).
10. Martin, F.-P. J. *et al. Mol. Syst. Biol.* **4,** 157 (2008).

ELECTRONICS

Industry-compatible graphene transistors

An innovative technique has been developed to manufacture graphene transistors that operate at radio frequencies and low temperatures. The process brings the devices closer to applications. SEE LETTER P.74

FRANK SCHWIERZ

To an increasing extent, modern society relies on advances in wireless communications. The backbone of wireless systems is radiofrequency (RF) transistors that are able to amplify signals and provide electronic gain at high frequencies. Unfortunately, these abilities degrade with increasing frequency, but emerging applications require ever higher operating frequencies. On page 74 of this issue, Wu *et al.*[1] describe transistors made from graphene — a carbon sheet just one atom thick — that hold promise for RF applications.

Two parameters are used to assess the frequency performance of an RF transistor: the cut-off frequency, f_T, at which the device's current gain drops to unity; and the maximum frequency of oscillation, f_{max}, at which the power gain becomes unity. One way to enhance the frequency performance of transistors is to use new materials that have high charge (carrier) mobility. The ultra-high mobilities observed[2,3] in graphene attracted the attention of device engineers immediately after their discovery, and intensive research[4,5] on RF graphene transistors is now under way.

Significant progress has been made since the demonstration[6] of the first gigahertz graphene transistors in 2008. Most notably, in February 2010, a group reported[7] a field-effect transistor (FET, the type of transistor most frequently used in electronics) made from graphene that broke the 100-GHz-f_T mark. And only a few months later, researchers demonstrated[8] a graphene FET that has an f_T of 300 GHz. Wu *et al.*[1] now report graphene FETs with gate electrodes of remarkably short length (40 nanometres) and f_T as high as 155 GHz. This result certainly does not represent a new record in frequency performance for RF transistors, and one might say that this is just another report on the good performance of graphene transistors. In fact, it is more than that in several respects.

Most groups make graphene by mechanical exfoliation, a method described[2] by Nobel prizewinners Konstantin Novoselov and Andre Geim. Mechanical exfoliation consists of peeling graphene flakes off a graphite crystal, and is a neat and practical method for university labs; the 300-GHz-f_T transistor mentioned above is made from exfoliated graphene. To make graphene attractive for the electronic-chip industry, however, reliable large-scale preparation schemes are needed. One such scheme, pioneered by Berger and de Heer[9], is the growth of graphene, by a method known as epitaxy, on silicon carbide wafers. A second option is to use a process known as chemical vapour deposition (CVD) to grow graphene on a metal, and then to transfer the graphene from the metal onto an insulating substrate, which most commonly consists of silicon with a top layer of silicon oxide (SiO_2)[10]. Wu and colleagues[1] now present a promising modification of the latter approach, which is to use a diamond-like carbon film as the top layer. Using this instead of SiO_2 is thought to result in better carrier transport in graphene FETs.

The authors[1] fabricated graphene transistors with gate lengths in the 40–550-nm range. They demonstrate reproducible measured characteristics for 30 devices and cut-off frequencies up to 155 GHz. Although this f_T value does not exceed that obtained previously[8], it is the highest f_T reported for CVD graphene transistors. Wu *et al.* also provide the first RF data for CVD graphene FETs on diamond-like carbon. With this work, another industry-compatible technology option for RF graphene FETs is now available.

What's more, Wu and colleagues are the first to investigate graphene FETs at very low temperatures. They show that the f_T of their transistors remains essentially constant between 300 kelvin and liquid-helium temperatures (4.2 kelvin), proving that graphene transistors could represent an alternative to conventional silicon- and III-V-semiconductor-based FETs for use at cryogenic temperatures, for example in space-based applications. It should be noted that proper operation of devices at 4.2 K is not a matter of course. There were serious concerns that carrier freeze-out might degrade the performance of silicon transistors of the MOSFET type at low temperatures. Fortunately, experiments[11] showed that this was not the case.

Finally, Wu and co-workers discuss not only the merits but also, and quite thoroughly, the problems of graphene transistors. The f_T performance of graphene FETs is known to be very competitive. For most applications, a high f_T is certainly desirable, but more important than a high f_T would be high power gain and f_{max}. Unfortunately, graphene FETs still suffer from low f_{max}. The 550-nm-gate graphene FETs of Wu *et al.* display an f_{max} of only 20 GHz. Although this is the highest f_{max} reported so far for graphene RF FETs, it is much lower than that of competing RF FETs (Fig. 1). Moreover, in

© 2011 Macmillan Publishers Limited. All rights reserved

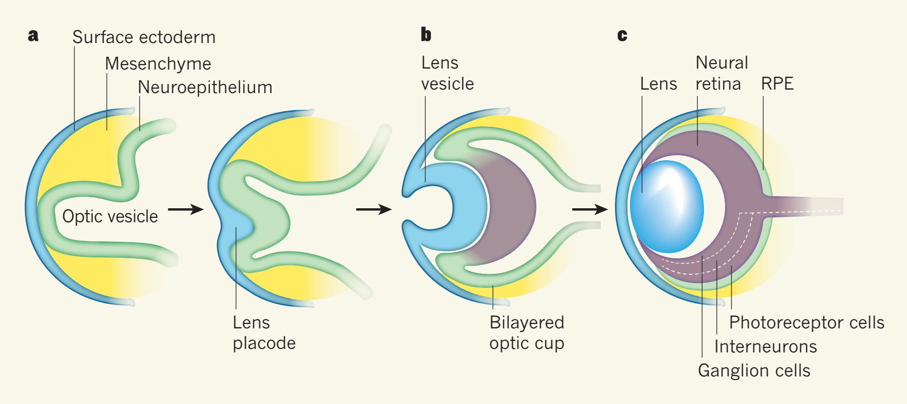


Figure 1 | Eye development. **a**, At early stages of eye development, the surface ectoderm thickens and invaginates together with the underlying neuroepithelium of the optic vesicle. **b**, The inner layer of the bilayered optic cup gives rise to neural retina and the outer layer gives rise to the retinal pigmented epithelium (RPE) (**c**). The mature neural retina (**c**) comprises three cellular layers: photoreceptors, interneurons (horizontal, amacrine and bipolar cells), and retinal ganglion cells. Eiraku *et al.*[1] generated optical cups *in vitro* from embryonic stem cells.

appropriately expressed the distinctive molecular markers of both the neural retina and the RPE, confirming their identity; another indicator was visible as RPE pigmentation.

An even more striking proof that these are genuine retinas is that, in culture, the synthetic optic cups undergo cell differentiation. Indeed, retinal progenitor cells — the multipotential cells of the neural retina — divided and differentiated into all the main retinal neuronal cell types, including photoreceptors. These events seem to follow the normal temporal sequence of retinal tissue formation, and the resulting cells were correctly organized in the appropriate cellular layer.

But even though optic cups can now be grown in culture from ES cells, we still don't fully understand the principles underlying their development. For instance, it is surprising that optic cups can form independently of any interaction of the neuroepithelial cells with surface ectoderm or mesenchymal tissue that would normally surround them in a developing embryo (Fig. 1). Eiraku *et al.*[1] propose that the ES-cell-derived retinal cells have a latent intrinsic order, and that collections of cells can self-pattern and undergo dynamic morphogenesis by obeying a sequential combination of local rules and internal forces within the epithelium.

However, Eiraku and colleagues' powerful *in vitro* system has great potential as it can be manipulated to define the molecular interactions that are essential for eye development. Moreover, if functional outer rod segments — where the protein complexes responsible for phototransduction are located — can be produced in longer-term cultures, this 3D system will be invaluable for functional studies examining the response of the retina to light.

What's more, development of an equivalent human 3D system could offer the prospect of disease modelling and drug testing using induced pluripotent stem cells generated from patients' tissues. Most forms of untreatable blindness result from the loss of photoreceptor cells, leaving other retinal neurons intact. In mice, transplantation of photoreceptor precursor cells isolated from the developing mouse retina can repair adult retinas[8]. A major challenge is to obtain sufficient numbers of photoreceptor precursors at the appropriate stage of development from a renewable cell source. This 3D system for culturing ES cells[1] may solve that problem by providing synthetic retinas at defined stages of development from which precursors can be isolated more readily for use in transplantation. ■

Robin R. Ali *is in the Department of Genetics, Institute of Ophthalmology, University College London, London EC1V 9EL, UK.*
Jane C. Sowden *is in the Developmental Biology Unit, Institute of Child Health, University College London, London WC1N 1EH, UK.*
e-mails: r.ali@ucl.ac.uk; j.sowden@ich.ucl.ac.uk

1. Eiraku, M. *et al. Nature* **472,** 51–56 (2011).
2. Zuber, M. E., Gestri, G., Viczian, A. S., Barsacchi, G. & Harris, W. A. *Development* **130,** 5155–5167 (2003).
3. Lamba, D. A., Karl, M. O., Ware, C. B. & Reh, T. A. *Proc. Natl Acad. Sci. USA* **103,** 12769–12774 (2006).
4. Osakada, F. *et al. Nature Biotechnol.* **26,** 215–224 (2008).
5. Yang, C. *et al. FASEB J.* **24,** 3274-3283 (2010).
6. Meyer, J. S. *et al. Proc. Natl Acad. Sci. USA* **106,** 16698–16703 (2009).
7. Ikeda, H. *et al. Proc. Natl Acad. Sci. USA* **102,** 11331–11336 (2005).
8. MacLaren, R. E. *et al. Nature* **444,** 203–207 (2006).

OCEANOGRAPHY

When glacial giants roll over

The energy released by capsizing icebergs can be equal to that of small earthquakes — enough to create ocean waves of considerable magnitude. Should such 'glacial tsunamis' be added to the list of future global-warming hazards?

ANDERS LEVERMANN

About half of Greenland's annual ice loss occurs through solid-ice discharge; in Antarctica such calving processes account for almost all ice loss. The resulting icebergs come in various sizes and shapes, some several hundred metres high. Immediately after they break off, when their height exceeds their horizontal extent, these floating giants can be unstable and capsize. In a paper in the *Annals of Glaciology*, MacAyeal and colleagues[1] have estimated the energy that is released when icebergs roll over. They find that this can be as large as that of an earthquake of magnitude 5–6 on the Gutenberg–Richter scale, depending on the iceberg's dimensions.

As one of several possibilities, a proportion of this energy can generate a surface gravity wave — a tsunami. MacAyeal *et al.* provide a theoretical analysis of the potential of iceberg capsize to generate tsunamis. Assuming a simplified, but not completely unrealistic, rectangular geometry, they calculate the potential energy before and after capsizing. The difference is the energy released from the roll-over. Thin icebergs do not carry a lot of potential energy, whereas ice-cube-shaped icebergs, which are as thick as they are high, have the same potential energy before and after turning over. Thus the most energy is released by icebergs that are half as thick as they are high — and can be equivalent to the explosion of several thousand tonnes of TNT.

According to MacAyeal and colleagues[1], energy release increases with the fourth power of an iceberg's height (Box 1). But not all of

© 2011 Macmillan Publishers Limited. All rights reserved

NATURE

50 Years Ago

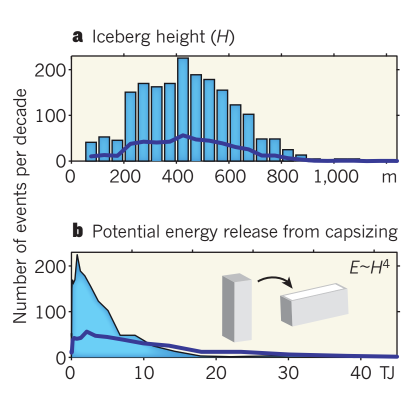
Hospital Infection: Causes and Prevention. A systematic approach to the causes and prevention of hospital infection is much to be welcomed. Accurate records are meagre and the problem is one which belongs to everybody and, consequently, to no one. Since streptococcal infections now cause no real difficulties—they still respond to penicillin and have acquired no resistance to the drug—the book is mainly concerned with the staphylococcal infections which, because of their resistant strains, are the main source of infection and worry in hospitals today... The text is clear and logically presented, and adds to the value of a book which should be useful not only to pathologists and bacteriologists but also to surgeons, paediatricians, sister tutors, hospital administrators, and equally important, hospital architects.
From *Nature* 8 April 1961

100 Years Ago

Dr. A. C. Johansen gives a summary account of the recent investigations on plaice and plaice fisheries in Danish waters... It includes an account both of the market statistics of plaice landed and of the special scientific investigations and experiments which have been carried out. The market conditions in Denmark are exceptional ... the chief demand is for fish that are landed alive... [As] there is a size limit (25.6 cm.) below which they are not allowed to be landed, and the fish under this size are returned to the sea, the actual destruction of small fish is insignificant. It appears that since the introduction of the size limit the Danish plaice fisheries in the North Sea have increased, and the report speaks in favour of an international size limit for plaice for all countries carrying on fisheries in the North Sea.
From *Nature* 6 April 1911

BOX 1

Iceberg height and capsize energy



These results come from simulations for the Antarctic made at my institute with the Potsdam Parallel Ice Sheet Model[3,4]. **a,** Frequency distribution of iceberg height, *H*, in discharge events per decade, assuming a quadratic ground area proportional to H^2. Iceberg discharge is computed from the vertical extent of the ice sheet and its velocity distribution in the present-day equilibrium state. The results show a peak in the abundance of icebergs with a height of around 400 metres. This differs from observations of icebergs that are freely floating in the ocean[5,6], and is probably due mainly to the assumption that icebergs break off only in whole slices. But it provides an indication of how much discharge occurs at various heights of the ice-sheet margin at which iceberg calving occurs. **b,** Potential energy, *E*, released from capsizing of these icebergs, with $E \sim H^4$. The shading depicts results for an aspect ratio (thickness/height) of ¼. Maximum energy is released for an aspect ratio of ½ (thick blue lines in both **a** and **b**). 1 TJ = 10^{12} J. For comparison, 4.2 TJ is the energy released by the explosion of one kilotonne of TNT. **A.L.**

this energy is available for tsunami generation. The authors suggest at least five other ways in which energy can be dissipated, ranging from the small rocking motion of the capsized icebergs themselves to mesoscale turbulent friction within the ocean. On the basis of a scaling analysis and by analogy with submarine landslides, however, the authors propose that several per cent of this energy can be translated into a tsunami wave. This is the crucial and most complicated part of the problem.

Once the energy transfer is known, the question arises of what height of tsunami is produced by an iceberg capsizing. The authors find that in typical iceberg regions located off the coast, but not yet over the deep ocean, the tsunami crest can reach up to 1% of the initial iceberg height — that is, 4 metres for an average iceberg from Antarctica (Box 1) but possibly up to 10 metres for the tallest icebergs on Earth. These numbers are comparable to the open-ocean crest heights of the devastating tsunami in the Indian Ocean in 2006 and the recent event in Japan. But they need to be understood as simple estimates — as ballpark values that show that capsizing icebergs may cause considerable tsunami waves. MacAyeal and colleagues provide a clean and beautifully simple theoretical framework for further studies of the subject. As stated by the authors, laboratory experiments, field observations and model simulations are essential to better understand the phenomenon.

Tsunamis generated by sudden iceberg motion have been reported to cause severe but localized damage in some Greenland fjords, where harbours have been destroyed by the wave[2]. Whether they pose a threat to more populated areas remote from their point of origin merits investigation. In principle, tsunamis pass across the deep ocean with practically no dissipation because friction there is low, and they are hardly disturbed by ocean currents such as the Antarctic Circumpolar Current or the North Atlantic Current.

We will need a great range of scientific insights — from iceberg-calving physics to wave generation from sudden iceberg motion — before we can say whether such glacial tsunamis will become more abundant in a world experiencing global warming. In principle, it is possible that iceberg-generated tsunamis could travel across the oceans and reach areas more populated than Antarctica. But as MacAyeal and colleagues speculate[1], even in the south polar region, explosive energy release might have wider effects by causing the collapse of floating ice shelves, thereby influencing global sea-level rise. ■

Anders Levermann *is at the Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany.*
e-mail: anders.levermann@pik-potsdam.de

1. MacAyeal, D. R., Abbot, D. S. & Sergienko, O. V. *Ann. Glaciol.* **52** (58), 51–56 (2011).
2. www.youtube.com/watch?v=_2NvwInKVtU
3. Winkelmann, R. *et al. Cryosphere Discuss.* **4,** 1277–1306 (2010).
4. Martin, M. A. *et al. Cryosphere Discuss.* **4,** 1307–1341 (2010).
5. Hamley, T. C. & Budd, W. F. *J. Glaciol.* **32,** 242–251 (1986).
6. Jacka, T. H. & Giles, A. B. *J. Glaciol.* **53,** 341–356 (2007).

© 2011 Macmillan Publishers Limited. All rights reserved

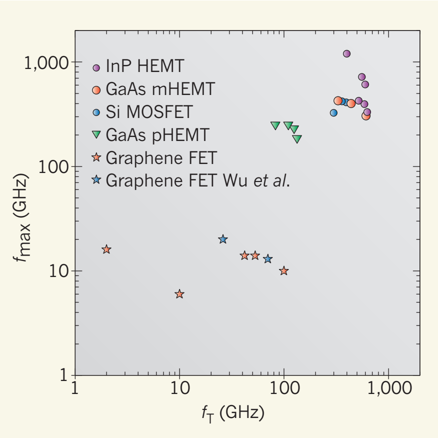


Figure 1 | Frequency performance of graphene transistors. Maximum frequency of oscillation, f_{max}, versus cut-off frequency, f_T, for graphene field-effect transistors (FETs) and competing radiofrequency FETs: indium phosphide high electron mobility transistor (InP HEMT), gallium arsenide metamorphic HEMT (GaAs mHEMT), silicon metal-oxide-semiconductor FET (Si MOSFET), and GaAs pseudomorphic HEMT (GaAs pHEMT). The red stars designate FETs made from 'epitaxial' graphene, whereas blue stars denote Wu and colleagues'[1] FETs, which were made from graphene grown by chemical vapour deposition.

conventional RF FETs, f_{max} commonly improves with shorter gate lengths, but the opposite is the case for Wu and colleagues' graphene FETs.

The main reason for the disappointing f_{max} is the unsatisfying, weak saturation of the device's drain current. Experience with conventional RF FETs clearly shows that, to exploit their full frequency potential, FETs need to be operated in a regime of strong current saturation. One explanation for the weak saturation in graphene FETs is the high electrical resistance between the device's electrodes (source and drain) and its graphene channel[12]. Unfortunately, a reliable way of significantly reducing such contact resistance in graphene devices is still lacking. Another issue that affects current saturation is the fact that graphene lacks a bandgap (an energy range where no electron states can exist). The huge gap between the f_{max} performance of graphene FETs and that of competing silicon and III-V FETs indicates that achieving strong current saturation and low contact resistance is crucial to making graphene RF FETs more competitive, and to open the door to their application in electronic circuitry. Although closing the gap seems hardly possible at the moment, we should remain optimistic and keep in mind the short history of graphene RF transistors and the huge progress made in the field since 2008. ■

Frank Schwierz *is at the Technische Universität Ilmenau, Postfach 100565, 98684 Ilmenau, Germany.*
e-mail: frank.schwierz@tu-ilmenau.de

1. Wu, Y. *et al. Nature* **472,** 74–78 (2011).
2. Novoselov, K. S. *et al. Science* **306,** 666–669 (2004).
3. Morozov, S. V. *et al. Phys. Rev. Lett.* **100,** 016602 (2008).
4. Avouris, Ph. *Nano Lett.* **10,** 4285–4294 (2010).
5. Schwierz, F. *Nature Nanotechnol.* **5,** 487–496 (2010).
6. Meric, I. *et al. Tech. Dig. IEDM*, paper 21.2 (2008).
7. Lin, Y.-M. *et al. Science* **327,** 662 (2010).
8. Liao, L. *et al. Nature* **467,** 305–308 (2010).
9. Berger, C. *et al. J. Phys. Chem. B* **108,** 19912–19916 (2004).
10. Li, X. S. *et al. Science* **324,** 1312–1314 (2009).
11. Ekanayake, S. R., Lehmann, T., Dzurak, A. S., Clark, R. G. & Brawley, A. *IEEE Trans. Electron Devices* **57,** 539–547 (2010).
12. Wu, Y. Q. *et al. Tech. Dig. IEDM* 226–228 (2010).

REGENERATIVE MEDICINE

DIY eye

Generation of complex organs *in vitro* is a major challenge in regenerative medicine. But it is not an impossible one: an entire synthetic retina has now been generated from embryonic stem cells. SEE ARTICLE P.51

ROBIN R. ALI & JANE C. SOWDEN

In this issue, Eiraku *et al.*[1] provide a series of extraordinary videos recording the formation of an embryonic mouse eye: for the first time, we see unfolding in real time the beautiful events that shape the early stages of mammalian eye development. But even more remarkable is that these are not recordings from live animals, but of self-organizing three-dimensional (3D) cultures of embryonic stem cells.

By the sixth week of human development, the rudiments of the mature eye are visible: bilayered optic cups, partially encapsulating the lens vesicles, have formed from the eye-field region of the anterior neural plate and the overlying surface ectoderm (Fig. 1). From the inner layer of the cup, the complex laminar structure of the neural retina will develop, with light-sensing photoreceptor cells connecting through interneurons to the retinal ganglion cells whose axonal processes project to the higher visual centres in the brain.

Elucidation of the mechanisms underlying embryonic eye development began more than a century ago. In one of his most significant experiments, Hans Spemann, a founder of developmental biology, showed that if the optic vesicle (the structure that eventually evolves into the optic cup) is destroyed, the lens fails to form. The interaction of the surface ectoderm (from which the lens derives) with the underlying optic vesicle has been considered a classical example of embryonic induction — the process by which one cell group signals to a neighbouring group and influences their future development. An array of genes has now been identified, many of which encode transcription factors or growth factors that are essential for the formation of the optic cup.

The likelihood of growing a complex organ such as an eye in a dish, however, has seemed remote and futuristic, although this distant frontier of regenerative medicine constantly moves closer. In the past decade, inspiring work[2] has shown that expression of eye-field transcription factors can lead to eye formation in unusual locations along the body of *Xenopus* frogs. Moreover, following the generation of human embryonic stem (ES) cells, it has proved possible[3,4] to direct their differentiation towards the retinal lineage and generate both retinal pigmented epithelium (RPE) and retinal neurons (Fig. 1). Cell-culture approaches have mainly sought to maximize the development of specific cell types with the potential aim of transplanting such cells for therapeutic purposes.

In vitro, RPE cells derived from ES cells self-organize into a characteristic simple monolayer. By contrast, reproducing the more complex and precise laminar organization of the neural retina presents a difficult tissue-engineering challenge. But reports describing lens-like structures[5] and retinal progenitor rosettes in ES-cell cultures[6] hinted at some potential for organization of eye tissue *in vitro*.

Now, Eiraku *et al.*[1] (page 51) reveal with startling beauty and remarkable clarity that the complex process of evagination of the optic vesicle, and then its invagination to form the bilayered cup, can occur spontaneously in culture, starting with a population of homogeneous pluripotent cells — cells that can differentiate into any cell type (see Fig. 1 of the paper[1] and the supplementary videos).

The key to this advance was that Eiraku and colleagues did not just simplify their previous[7] differentiation protocol for ES cultures, but also added Matrigel, which includes extracellular-matrix components. Under these conditions, and using a green fluorescent protein (GFP) reporter gene expressed in the eye field and the neural retina, they found that a neuro-epithelium-like layer of GFP-positive cells evaginated from the sides of hollow balls of ES cells, in a process reminiscent of optic-vesicle formation. Over time, the optic vesicles spontaneously underwent dynamic morphogenesis and formed bilayered cups. The cups

© 2011 Macmillan Publishers Limited. All rights reserved

ECOLOGY

Diversity favours productivity

A consequence of Darwin's 'principle of divergence' is that loss of species can harm the functioning of ecosystems. A study of algal communities in artificial streams suggests that he was right. SEE LETTER P.86

ANDY HECTOR

'Could do better!' This was last year's disappointing report on attempts to meet the Convention on Biological Diversity's goal of slowing the rate of global biodiversity loss, in an assessment for the 2010 United Nations International Year of Biodiversity[1]. The outlook for this century suggests that we're set for a biodiversity crisis that may, within a few centuries, join the ranks of the previous 'Big Five' mass extinctions[2]. Many ecologists are busy trying to work out what this could mean for the ecosystem services that we benefit from but largely take for granted, and on page 86 of this issue, Cardinale[3] presents one of the latest advances in this area. Using laboratory experiments with artificial stream ecosystems, he has shown that complex freshwater habitats require diverse communities of algal species, occupying different environmental niches, in order to be productive and to maintain water quality.

Human activities now fix more nitrogen into the biosphere than all natural processes combined. Eventually, through the run-off of fertilizers from agricultural fields and similar processes, much of this nitrogen and other nutrients will end up in our freshwater and coastal environments. This excessive nutrient loading can have negative impacts both on the health of humans and on the environment, and necessitates costly clean-up mechanisms. It is against this backdrop that Cardinale asks what part algal diversity plays in maintaining water quality by taking up nitrogen.

Cardinale used a high-tech set-up of artificial streams, which varied in the complexity of their physical environments in terms of frequency of disturbance and by having variable or constant stream-flow velocities. This set-up generated experimental ecosystems that had a relatively high or low number of different environmental niches. He inoculated these experimental streams with algal communities of varying diversity, and studied how productive the resulting model ecosystems were and how much nitrogen they captured from the water column.

His results provide some of the strongest support so far for a hypothesis that dates back to Charles Darwin. In his 'principle of divergence'[4], Darwin proposed that species evolve into different niches through adaptation to different environmental conditions. He thought that this should lead to the evolution of communities of complementary species and an ecological "division of labour" — by analogy with the economic division of labour noted in pin manufacturing by Adam Smith — that can increase overall resource capture and productivity[5,6]. As a consequence, biodiversity loss can have negative effects on ecosystem functioning by leaving ecological niches vacant or underused. Support for this idea was provided recently in a study[7] of marine microbes that were experimentally evolved into either specialists (which exploit narrow environmental niches) or generalists (which thrive under a broader range of conditions). The relationship between biodiversity and ecosystem functioning was found to be stronger for the communities of specialists tightly adapted to particular niches.

Cardinale's research[3] provides further evidence of the detrimental effect of biodiversity loss on ecosystem functioning. In his complex stream environments, only a diverse community of complementary algal species could exploit all the available niches (see Fig. 2 of the paper[3]). When some species were absent, their ecological niches were left vacant or underused because their absence could not be fully compensated for by competitors adapted to other niches. The presence of vacant or underused niches led in turn to reduced nitrogen uptake and decreased productivity of communities.

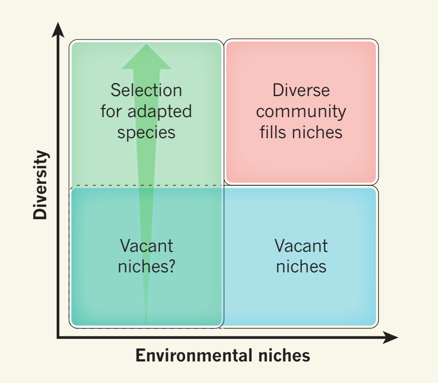


Figure 1 | Complex environments require a diverse community of complementary species. Cardinale[3] has manipulated the diversity (number of algal species) and the number of environmental niches (complexity) in artificial stream ecosystems. He reports that complex habitats require a diverse community of complementary species to fully use the available niches (top right). Loss of biodiversity can result in vacant or underused niches and reductions in resource capture and productivity (bottom right). When there are few niches (left-hand side), environmental filtering selects species adapted for those niches; the greater the number of species, the greater the selection (arrow). Vacant niches could also occur in non-complex systems if adapted species are not present (bottom left).

Identifying the detailed biological mechanisms responsible for the effects of biodiversity on ecosystem functioning has proved difficult, particularly for complex communities of plants and their bacterial and fungal partners and adversaries. Most studies so far have therefore had to settle for identifying general classes of proximate mechanisms using a variety of statistical approaches[8,9], as indeed Cardinale did in his study[3] to separate species complementarity and selection effects. But Cardinale's experimental system was able to address biological mechanisms more directly than most previous studies, for two reasons. First, the preferred environmental niches of the algal study species were already quite well defined. Such definition is sorely needed in other investigations of biodiversity and ecosystem functioning, and in community ecology in general. Second, the number of environmental niches could be changed directly by creating stream ecosystems of varying complexity.

The combination of this experimental manipulation of habitat-niche complexity with the more customary manipulation of community diversity allowed Cardinale to provide an unusually direct demonstration of how the two factors interact (Fig. 1). In simple habitats, he observed that a low diversity of species is sufficient to occupy all of the available niches as long as appropriately adapted species are present. But complex habitats require diverse communities of complementary species to use the environment more fully, as proposed by Darwin's principle of divergence. When Cardinale simulated rapid (in evolutionary terms) species loss by using low-diversity communities, this led to vacant or underused niches in the complex stream environments.

Critics of the idea that biodiversity influences ecosystem functioning, and of Cardinale's experiments, may argue that our current and future homogenized environments might function adequately with low biodiversity, as long as the species that are left are adapted to the prevailing conditions. The assumption that we will be left with low-diversity communities of species that function well is risky, but may sometimes be true, at least if we ignore such considerations as the stability of ecosystem functioning over time[10] and the need for ecosystems to perform multiple functions[11]. For example, although Cardinale's simple stream environments could host only low-diversity algal communities, these

© 2011 Macmillan Publishers Limited. All rights reserved

ecosystems were roughly as productive, and captured as much nitrogen, as more complex stream habitats that supported a wider range of species. In fact, in the simple environments, a monoculture of one species was the most productive of all (compare the top and bottom rows of Fig. 1a and 1b in the paper[3]). This suggests that low-diversity algal communities may be adequate to support a small number of functions in the homogeneous, managed freshwater systems that currently characterize much of our world, at least over the short term. But Cardinale's study also shows that such communities may not be capable of fully using all the available niches if we progress to a post-Anthropocene world containing more varied and complex environments, similar to the natural systems where these species evolved. ■

Andy Hector *is at the Institute for Evolutionary Biology and Environmental Studies, University of Zurich, CH-8057 Zurich, Switzerland.*
e-mail: andrew.hector@uzh.ch

1. Butchart, S. H. M. *et al. Science* **328,** 1164–1168 (2010).
2. Barnosky, A. D. *et al. Nature* **471,** 51–57 (2011).
3. Cardinale, B. J. *Nature* **472,** 86–89 (2011).
4. Darwin, C. *Charles Darwin's Natural Selection: Being the Second Part of his Big Species Book Written from 1856 to 1858* (ed. Stauffer, R. C.) (Cambridge Univ. Press, 1975).
5. Hector, A. in *Darwin und die Botanik* (eds Stöcklin, J. & Höxtermann, E.) 182–191 (Basilisken, 2009).
6. Hector, A. & Hooper, R. *Science* **295,** 639–640 (2002).
7. Gravel, D. *et al. Nature* **469,** 89–92 (2011).
8. Loreau, M. & Hector, A. *Nature* **412,** 72–76 (2001).
9. Hector, A. *et al.* in *Biodiversity, Ecosystem Functioning, and Human Wellbeing* (eds Naeem, S. *et al.*) 94–104 (Oxford Univ. Press, 2009).
10 Hector, A. *et al. Ecology* **91,** 2213–2220 (2010).
11. Hector, A. & Bagchi, R. *Nature* **448,** 188–190 (2007).

REPRODUCTIVE BIOLOGY

Bone returns the favour

There are well-established links between the reproductive system, metabolism and skeletal growth. But it comes as a surprise that the skeleton — more specifically, the bone hormone osteocalcin — modulates fertility.

SONYA M. SCHUH-HUERTA & RENEE A. REIJO PERA

Fertility is a complex process: it is regulated by numerous genetic and environmental factors, and involves coordinate regulation among several organ systems. It is also directly associated with the quantity and quality of sperm and eggs[1], and is affected by disease and ageing. Reporting in *Cell*, Oury *et al.*[2] identify a new player. Osteocalcin — a hormone secreted by bone — regulates male fertility by increasing testosterone levels.

The sex hormones testosterone and oestrogen have central roles in reproduction. Testosterone promotes the development and function of the testes and stimulates sperm production and survival[3]. Oestrogen affects ovarian function and promotes egg maturation and ovulation. Both hormones also induce the development of secondary sex characteristics — in males, for example, the deepening of the voice, patterns of facial and body hair, and larger body size, and, in females, the development of breasts and hips[4].

As central regulators of reproduction, testosterone and oestrogen affect bone, muscle, fat, sexual function, mood and cognition. One essential function of these hormones is the regulation of both skeletal growth and accumulation of bone mass — as seen in the growth spurts of puberty, a time during which sex-hormone levels rise markedly[5]. With age, or in diseases that cause gonadal dysfunction, the decline or loss of oestrogen and testosterone is accompanied by a decline in skeletal mass[6,7]. But does the skeletal system have a reciprocal effect on the reproductive system? Yes, according to Oury and colleagues[2], who demonstrate that osteocalcin, which is secreted by bone-forming osteoblast cells, plays a crucial part in stimulating testosterone production by the testis.

The authors cultured factors derived from bone, fat, muscle or skin together with mouse testes. Specifically, osteoblast-derived factors increased testicular production of testosterone. This effect was limited to the male: osteoblasts did not stimulate testosterone or oestrogen production by the ovaries. The responsive cells were Leydig cells, which secrete testosterone in the intact testis.

In addition to being a major hormone that osteoblasts secrete, osteocalcin is also a marker of bone formation and has roles in energy metabolism and glucose homeostasis[8]. Male mice lacking the gene that encodes osteocalcin show metabolic and skeletal alterations[9], are fat and are poor breeders. Oury and co-workers therefore postulated that osteocalcin may be the main bone hormone that increases testosterone production by Leydig cells. Indeed, they find that osteocalcin leads to dose-dependent increases in testosterone; that only osteoblasts that secrete osteocalcin produce this effect; and that osteocalcin injection into male mice increases circulating levels of testosterone.

The researchers also investigated osteocalcin's role in fertility and found that male mice deficient in this hormone have greatly reduced testosterone levels, smaller reproductive organs, lower sperm counts, fewer sperm in advanced stages of development, and a greater number of dying sperm in their testes. Moreover, the animals fathered fewer offspring. By contrast, male mice lacking the gene that inhibits osteocalcin activity (a gain in osteocalcin function) had slightly enhanced fertility. The receptor that mediates osteocalcin's action is probably the G-protein-coupled receptor GPRC6A, and the authors also identify many components of the associated signalling pathway.

These findings reveal a previously unknown link between the reproductive and skeletal systems (Fig. 1), which does make sense. Bone is a highly dynamic tissue involved in whole-body energy metabolism and in maintaining calcium levels[5]. In many ways, the health of various organ systems is reflected in reproductive function and the ability to reproduce. And unless the body has sufficient energy stores, why would it invest energy in reproduction?

Several metabolic signals may act as gatekeepers of reproduction. For example, the timing of puberty is associated with nutritional status, weight and height[10]. Genome-wide association studies[11] have even identified specific genes that correlate with both the events of puberty and skeletal growth. Moreover, poor nutrition or extreme physical activity can delay puberty or cause loss of menstrual cycles and infertility[10]. Conversely, obesity is often linked with diminished fertility, causing lower sperm counts and a reduced chance of pregnancy[12]. The hormones leptin (produced by fat cells) and insulin (involved in glucose metabolism) are two primary signals that influence both body-weight regulation and reproductive function in humans and animals[10,12,13]. So it is likely that similar skeletal signals mediate physiological regulation of reproduction. Osteocalcin may be one such signal.

Osteocalcin may not exert its effects solely on sperm production and survival. Greatly diminished testosterone levels also reduce libido and reproductive function in mice and men[12,14]. For instance, Oury and colleagues' osteocalcin-deficient male mice[2] are fertile — despite having reduced sperm numbers — but produce fewer litters. These animals therefore probably have reduced mating behaviour. The animals also show other hormonal and metabolic changes, such as increased levels of oestrogen and luteinizing hormone, which is secreted by the brain and enhances testosterone levels in males. So osteocalcin may also influence fertility through testosterone's modulation of sexual behaviour in the brain and alteration of general reproductive and metabolic hormone profiles.

Does osteocalcin function similarly in

© 2011 Macmillan Publishers Limited. All rights reserved

massive than a typical present-day star such as our Sun. The first dense gaseous structures in which stars formed were very small; they had a total mass of the order of a million solar masses — approximately a million times smaller than a typical present-day galaxy such as the Milky Way. Molecular hydrogen, which formed efficiently in these dense regions, allowed the gas to radiate efficiently and lose its pressure support, making it collapse further[2]. Three-dimensional simulations[3,4] revealed that these structures form at the intersections of thin filaments, forming a cosmic web-like structure. They have also shown that a small fraction of the gas — of about 100 solar masses — flows along the filaments coherently towards the central region of each such dense knot, without any sign of fragmenting. However, recent simulations[5–8], of higher resolution than the earlier ones[3,4], have suggested that the gas in the central regions does eventually fragment into two or more distinct clumps, raising the possibility that the first stars formed in pairs, or in even higher-multiple systems.

Why would the companionship of the first stars matter for the rest of the Universe? As Mirabel and colleagues argue[1], a natural outcome of the latter hypothesis is for one member of a pair of massive stars to implode, leaving behind a black hole that remains gravitationally bound to its massive partner. The black hole could then pull material off the surface of its partner, and swallow it up. While devouring its partner, the black hole would return a fraction of the ingested energy in the form of copious amounts of X-rays. In fact, there are compelling examples of such 'micro-quasars' in the local Universe. They seem to be more common in smaller galaxies, as well as in galaxies whose chemical composition is closer to that of the pristine plasma of hydrogen and helium in the early Universe, unpolluted by the heavier atoms produced by subsequent generations of stars. The extrapolation of these local observations suggests that such binaries were more common in the earliest, small and primitive micro-galaxies.

If most of the first stars formed such binaries, they could have produced sufficient X-rays to significantly change the prevailing Swiss-cheese scenario. This possibility has been raised in the past[9–12], but is now worth considering more seriously in light of the new theoretical and observational evidence. Unlike the ultraviolet ionizing radiation from normal stars, X-rays with the right energy — of the order of 1 kiloelectronvolt — could travel across vast distances in the early Universe, ionizing and heating the plasma much more uniformly. If the X-rays were sufficiently prevalent, they would have a range of other interesting effects. For example, the extra heating would raise the pressure of the plasma everywhere, making it resistant to clumping, and more difficult to compress to form new galaxies[9].

On the other hand, X-rays could penetrate the successfully collapsing galaxies and ionize hydrogen and helium atoms in their interior. This would catalyse the formation of molecular hydrogen, and help the gas to cool and form new stars[13]. Such effects would leave their signatures in the spatial distribution of neutral and ionized hydrogen and helium in the Universe. These distributions might be mapped by measuring the 21-centimetre-wavelength radio emission from neutral hydrogen[14], and the scattering of cosmic microwave background radiation (relic radiation from the Big Bang) by free electrons, or by examining the absorption spectra of distant galaxies. Such measurements will be feasible for forthcoming experiments, and form a major goal of modern cosmology.

There are other possible sources of X-rays connected to the formation of the first stars, for example gas accretion onto the black-hole remnants left behind by the collapse of single stars[15,16]. Another possible source is supernovae: if the first stars exploded as supernovae, similar X-rays would be produced by thermal emission from the gas heated by these explosions, and by the collisions between the energetic electrons produced in the supernova explosion and the photons of the cosmic microwave background[9]. However, if micro-quasars were indeed as common, and as efficient producers of X-ray radiation, as Mirabel and colleagues argue[1], they may well have dominated X-ray production in the transition epoch, when the first stars started to shine in the Universe. They would then have been responsible for ending the dark ages in a smooth fashion. The hardest X-ray photons (those with energies above a few kiloelectronvolts) would be reaching Earth now, forming a feeble X-ray background. Existing measurements place an upper limit on the present-day value of this background that is consistent with this hypothesis[17]. The possibility of X-ray production by binary stars should prompt further theoretical modelling of the population of such binaries, including their abundance, radiation output and spectra, as well as modelling of the possible observable signatures they left behind. ■

Zoltán Haiman *is in the Department of Astronomy, Columbia University, New York, New York 10027, USA.*
e-mail: zoltan@astro.columbia.edu

1. Mirabel, I. F., Dijkstra, M., Laurent, P., Loeb, A. & Pritchard, J. R. *Astron. Astrophys.* **528,** A149 (2011).
2. Haiman, Z., Thoul, A. & Loeb, A. *Astrophys. J.* **464,** 523–538 (1996).
3. Bromm, V., Yoshida, N., Hernquist, L. & McKee, C. F. *Nature* **459,** 49–54 (2009).
4. Abel, T., Bryan, G. L. & Norman, M. L. *Science* **295,** 93–98 (2002).
5. Turk, M. J., Abel, T. & O'Shea, B. *Science* **325,** 601–605 (2009).
6. Stacy, A., Greif, T. H. & Bromm, V. *Mon. Not. R. Astron. Soc.* **403,** 45–60 (2010).
7. Prieto, J., Padoan, P., Jimenez, R. & Infante, L. Preprint at http://arxiv.org/abs/1101.5163 (2011).
8. Greif, T. *et al.* Preprint at http://arxiv.org/abs/1101.5491 (2011).
9. Oh, S. P. *Astrophys. J.* **553,** 499–512 (2001).
10. Venkatesan, A., Giroux, M. L. & Shull, J. M. *Astrophys. J.* **563,** 1–8 (2001).
11. Glover, S. C. O. & Brand, P. W. J. L. *Mon. Not. R. Astron. Soc.* **340,** 210–226 (2003).
12. Chen, X. & Miralda-Escudé, J. **602,** *Astrophys. J.* 1–11 (2004).
13. Haiman, Z., Abel, T. & Rees, M. J. *Astrophys. J.* **534,** 11–24 (2000).
14. Furlanetto, S. R., Oh, S. P. & Briggs, F. H. *Phys. Rep.* **433,** 181–301 (2006).
15. Madau, P. *et al. Astrophys. J.* **604,** 484–494 (2004).
16. Ricotti, M., Ostriker, J. P. & Gnedin, N. Y. *Mon. Not. R. Astron. Soc.* **357,** 207–219 (2005).
17. Dijkstra, M., Haiman, Z. & Loeb, A. *Astrophys. J.* **613,** 646–654 (2004).

EARTH SCIENCE

A new mechanical model for Tibet

A three-dimensional mechanical model of the Tibetan crust explains both the first-order features of GPS surface velocities and the contrast in the types of earthquake between northern and southern Tibet. SEE LETTER P.79

JEFFREY T. FREYMUELLER

The collision of India with Eurasia has built the Tibetan Plateau[1], by far the largest area of high topography on the planet, as the Indian plate has thrust beneath Tibet. The crust of the plateau is approximately twice the typical thickness, and great debates have raged over its development and mechanical properties. Temperatures are high within the thickened crust[2,3], which implies significant weakness of the middle and lower crust[4,5]. Several groups have proposed that the middle crust of Tibet is fluid enough to decouple the upper crust from the underthrusting Indian lithosphere, such that Tibetan middle crust might be extruded through a low-viscosity channel flow from beneath the high plateau either southward[4] or eastward[5]. But others have argued[6] that deformation at all depths

© 2011 Macmillan Publishers Limited. All rights reserved

in the Tibetan crust is coherent. On page 79, Copley *et al.*[7] use a three-dimensional (3D) mechanical model to explain the pattern of faulting and surface deformation in Tibet, and conclude that mechanical coupling between underthrust India and Tibet must be strong in southern Tibet. This is inconsistent with channel-flow models for southern Tibet[4].

The Tibetan Plateau deforms pervasively, as shown by Global Positioning System (GPS) measurements[8,9], and earthquakes in the brittle upper crust occur over the entire plateau[7]. Thrust faulting, indicative of shortening and crustal thickening, is found only on the margins of the plateau. In the southern interior, the earthquakes occur almost entirely on normal faults, indicating extensional stress, whereas in the northern interior strike–slip faulting dominates.

This contrast in the type of faulting between the northern and southern interior of Tibet has long been noted, but cannot be explained by changes in stresses induced by topography or plate motions alone. Despite the change in style of faulting, the surface velocities obtained from GPS measurements at sites across the entire plateau show substantial and nearly uniform strain between the major strike–slip faults, specifically contraction in the direction of plate convergence and extension orthogonal to it[10]. A successful mechanical model needs to explain both the difference in the style of faulting between the northern and southern regions and the relatively uniform surface strain field shown by the GPS data.

Copley *et al.*[7] approximate the Tibetan crust using a 3D viscous model, with long-term (geological timescale) deformation driven by imposed velocity boundary conditions and topographically induced stresses. The model predicts the stress field (internal forces) within the crust, which determines the style of faulting, and the motions of points on the surface, which can be compared to the GPS data. The authors tested different boundary conditions, reflecting vertical mechanical coupling or decoupling between the Tibetan crust and the underlying Indian lithosphere, which is assumed to be rigid.

The model with strong mechanical coupling best fits both the difference in the style of faulting and the GPS observations. In northern Tibet, which is not underlain by rigid Indian lithosphere, the model predicts a stress state that promotes strike–slip faulting, with surface deformation that reflects shortening in the direction of plate convergence and extension orthogonal to it. This deformation results from a combination of compression from plate motions and topographically induced stresses. In southern Tibet, where the crust is coupled to underthrust India, the vertical mechanical coupling induces an additional component of shear stress. This significantly changes the stress field so that normal faulting and east–west extension are favoured instead of strike–slip faulting (Fig. 1).

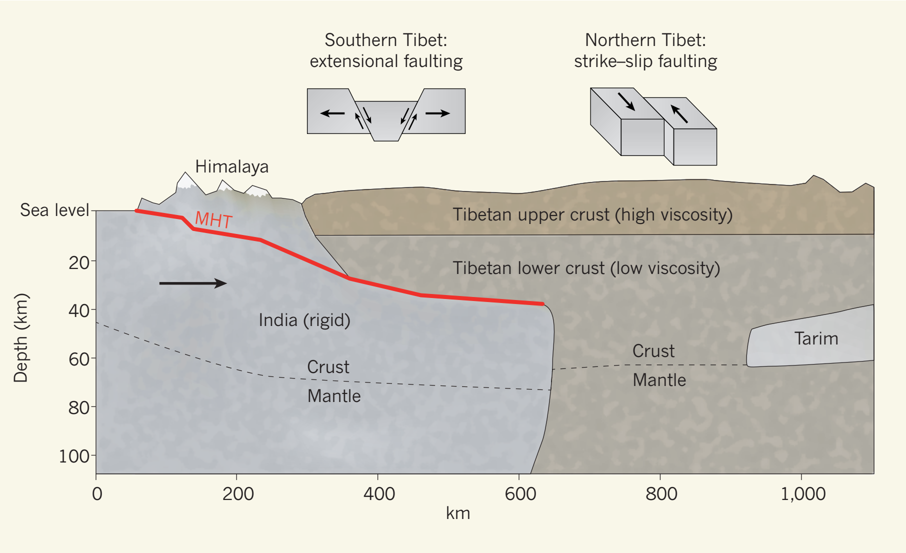


Figure 1 | Cross-sectional view of Tibet in the direction of plate convergence. The Indian lithosphere (crust and attached uppermost mantle) has been thrust beneath Tibet through millions of years of convergence between India and Eurasia. Dark-grey regions indicate rocks that originated with India; light grey indicates crust of the Tarim Basin, similarly underthrust from the north. The Main Himalayan Thrust (MHT) is the present main slip zone at seismogenic (earthquake-generating) depths, and the red line shows the deeper extension of the slip surface along the top of the Indian crust. Copley *et al.*[7] show that mechanical coupling between India and the overlying Tibetan crust explains the contrast between extensional faulting in southern Tibet and strike–slip faulting north of the limit of India (block diagrams). The weak (low viscosity) lower crust between India and the strong (high viscosity) upper crust flows in the direction of plate convergence, transmitting shear stress to the upper crust.

The authors[7] have succeeded in capturing the contrast between the observed faulting patterns in the northern and southern regions with a simple mechanical model, while also matching the first-order features of the observed GPS velocity field in northern Tibet. In southern Tibet, however, the GPS velocities contain a significant component due to the recoverable, elastic response to the build-up of stress on the Main Himalayan Thrust (MHT; Fig. 1) during the earthquake cycle. The shallow part of the MHT is frictionally locked, slipping mainly in large earthquakes. The presence of this locked region causes contraction normal to the Himalayan arc that extends a considerable distance into southern Tibet. With the next large earthquake, this region will spring back southwards, releasing centuries of stored strain energy in seconds to minutes. Together with an existing model for this elastic strain component[11], Copley *et al.*'s long-term deformation model can explain the first-order features of the GPS velocities across Tibet.

Their model does not match some details of the GPS velocity field. It underestimates the rate of east–west extension across the southern part of the plateau. In addition, it approximates Tibet as a continuous medium, and cannot include localized slip on the major strike–slip fault systems. In addition, the elastic model used for the strain from the MHT[11] does not work well for the region east of 90° E. This area near the eastern end of the Himalayan arc is complicated, and features a southward jump of the convergent front to the Shillong plateau in India. Further work will be needed to determine whether any of these factors simply reflect local variations or potential problems with the model.

Copley and colleagues[7] have placed important new constraints on the mechanical properties of Earth's continental lithosphere in its most extreme environment, and forced a critical evaluation of the channel-flow models for Tibet. Their model makes testable predictions of the average viscosity of the middle and lower crust in Tibet. The great debate is not finished, but it may have been channelled in a new direction. ■

Jeffrey T. Freymueller *is at the Geophysical Institute, University of Alaska Fairbanks, Fairbanks, Alaska 99775, USA.*
e-mail: jeff.freymueller@gi.alaska.edu

1. Royden, L. H., Burchfiel, B. C. & van der Hilst, R. D. *Science* **321,** 1054–1058 (2008).
2. Francheteau, J. *et al. Nature* **307,** 32–36 (1984).
3. Nelson, D. *et al. Science* **274,** 1684–1687 (1996).
4. Beaumont, C., Jamieson, R. A., Nguyen, M. H. & Lee, B. *Nature* **414,** 738–742 (2001).
5. Clark, M. K. & Royden, L. H. *Geology* **28,** 703–706 (2000).
6. England, P. & Molnar, P. *Science* **278,** 647–650 (1997).
7. Copley, A., Avouac, J.-F. & Wernicke, B. P. *Nature* **472,** 79–81 (2011).
8. Wang, Q. *et al. Science* **294,** 574–577 (2001).
9. Gan, W. *et al. J. Geophys. Res.* **112,** B08416 (2007).
10. Chen, Q. *et al. J. Geophys. Res.* **109,** B01403 (2004).
11. Bettinelli, P. *et al. J. Geodesy* **80,** 567–589 (2006).

© 2011 Macmillan Publishers Limited. All rights reserved

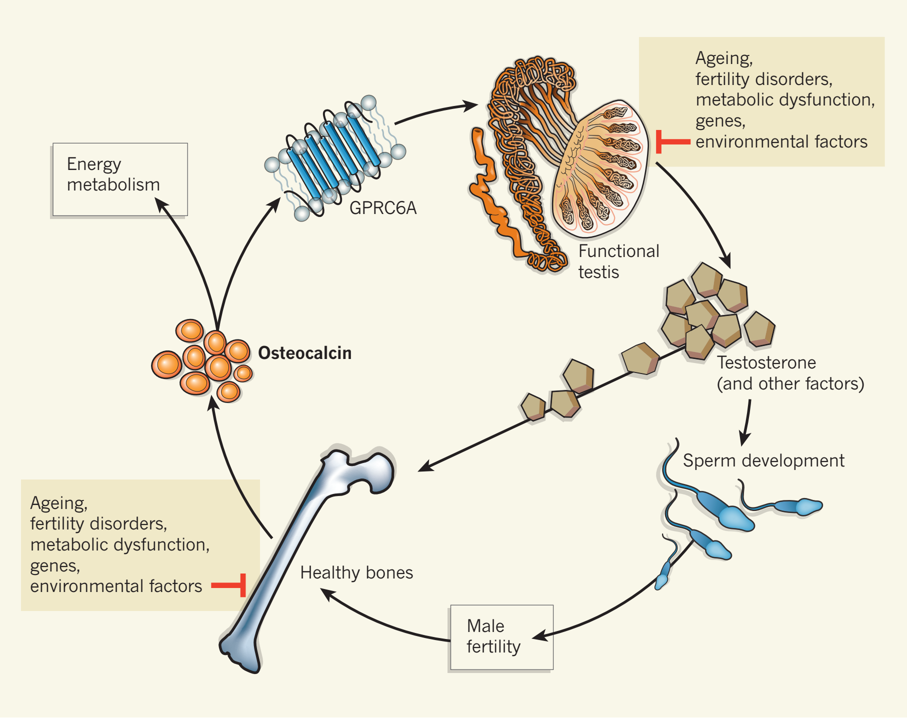

Energy metabolism
GPRC6A
Functional testis
Ageing, fertility disorders, metabolic dysfunction, genes, environmental factors
Osteocalcin
Testosterone (and other factors)
Sperm development
Ageing, fertility disorders, metabolic dysfunction, genes, environmental factors
Healthy bones
Male fertility

Figure 1 | A link between the skeleton and reproduction. Oury *et al.*[2] discover that the bone hormone osteocalcin, and its activity through the receptor GPRC6A, increases testosterone secretion by the mouse testis. Testosterone promotes sperm development, sexual function and fertility, as well as accrual of bone mass, bringing the connection between bone and the gonad full circle. The interaction between the skeleton and fertility underscores the importance of the health and function of several organ systems in the regulation of reproduction. Both systems are affected by a number of external factors including genes, metabolism and the environment.

humans? Male mice lacking this hormone, or its receptor, seem to represent a model for ageing men. In men, testosterone levels decrease with age, and although, with respect to testosterone, the levels of oestrogen and luteinizing hormone often rise, the latter cannot increase testosterone production[14]. There are also associated increases in body fat and decreases in bone mass, sperm counts and sexual function. Remarkably, osteocalcin-deficient mice also show similar features. Moreover, osteocalcin and its receptor — as well as other components of the signalling pathway — are present in human testes. All this evidence hints that osteocalcin might function similarly in men. So it is tempting to speculate that osteocalcin might be suitable to treat fertility defects and associated changes in ageing men. It is hoped that future work will reveal the role of this hormone and other metabolic factors in human reproduction, and provide new targets for contraceptives and treating infertility. ■

Sonya M. Schuh-Huerta *and* **Renee A. Reijo Pera** *are at the Institute for Stem Cell Biology and Regenerative Medicine, and in the Department of Obstetrics and Gynecology, Stanford University School of Medicine, Stanford, California 94305, USA.*
e-mails: sonyas1@stanford.edu; reneer@stanford.edu

1. Menken, J. & Larsen, U. *Ann. NY Acad. Sci.* **709,** 249–265 (1994).
2. Oury, F. *et al. Cell* **144,** 796–809 (2011).
3. Walker, W. H. *Steroids* **74,** 602–607 (2009).
4. Plant, T. M. & Witchel, S. F. in *Knobil and Neill's Physiology of Reproduction* 3rd edn (ed. Neill, J. D.) 2177–2230 (Elsevier, 2006).
5. Harada, S. & Rodan, G. A. *Nature* **423,** 349–355 (2003).
6. Manolagas, S. C., Kousteni, S. & Jilka, R. L. *Recent Prog. Horm. Res.* **57,** 385–409 (2002).
7. Khosla, S. *et al. J. Clin. Endocrinol. Metab.* **86,** 3555–3561 (2001).
8. Lee, N. K. *et al. Cell* **130,** 456–469 (2007).
9. Ducy, P. *et al. Nature* **382,** 448–452 (1996).
10. Ahima, R. S. *N. Engl. J. Med.* **351,** 959–962 (2004).
11. Ong, K. K. *et al. Nature Genet.* **41,** 729–733 (2009).
12. Mah, P. M. & Wittert, G. A. *Mol. Cell. Endocrinol.* **316,** 180–186 (2010).
13. Tena-Sempere, M. & Barreiro, M. L. *Mol. Cell. Endocrinol.* **188,** 9–13 (2002).
14. Kaufman, J. M. & Vermeulen, A. *Endocr. Rev.* **26,** 833–876 (2005).

COSMOLOGY

A smoother end to the dark ages

Independent lines of evidence suggest that the first stars, which ended the cosmic dark ages, came in pairs rather than singly. This could change the prevailing view that the early Universe had a Swiss-cheese-like appearance.

ZOLTÁN HAIMAN

After a spectacular birth, our Universe quickly became a dull place, with the glow of the Big Bang fading away and the first stars and galaxies yet to appear. These cosmic dark ages lasted for 100 million years. According to a growing body of evidence, the latest of which is described by Mirabel *et al.*[1] in a paper published in *Astronomy & Astrophysics*, many of the first stars that put an end to the dark ages may have formed in pairs.

The appearance of the first stars marked a significant milestone, separating the history of the Universe into two stages. The first stage is well understood: dark matter, primordial ionized plasma and radiation formed a nearly uniform mixture, expanding and cooling continuously with cosmic time. When the temperature of the plasma dropped below 3,000 kelvin, neutral hydrogen and helium atoms formed everywhere. Spatial variations in the density and temperature of the plasma were initially minuscule, but gravitational instability amplified these variations over time, and allowed dense gaseous structures to collapse in on themselves.

In the second stage, stars lit up inside these structures, and started wreaking havoc. The radiation of the stars penetrated the neutral cosmic plasma, once again ionizing and heating it, and modifying the formation of the subsequent generations of stars. At the time of the transition between the two stages, the 100-million-year-old Universe may have resembled Swiss cheese: cold and neutral background gas was filled with numerous, roughly spherical, hot, ionized holes surrounding the sites where the earliest stars had lit up. There is, however, another possibility, in which energetic X-ray radiation — not normally associated with stars — was present during the transition. The evidence that the first stars may have formed in pairs makes the latter hypothesis more likely.

Over the past decade, a theoretical paradigm has emerged in which the first stars formed in isolation and were about 100 times more

© 2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09941

Self-organizing optic-cup morphogenesis in three-dimensional culture

Mototsugu Eiraku[1,2], Nozomu Takata[1], Hiroki Ishibashi[3], Masako Kawada[1], Eriko Sakakura[1,2], Satoru Okuda[3], Kiyotoshi Sekiguchi[4], Taiji Adachi[3,5] & Yoshiki Sasai[1,2]

Balanced organogenesis requires the orchestration of multiple cellular interactions to create the collective cell behaviours that progressively shape developing tissues. It is currently unclear how individual, localized parts are able to coordinate with each other to develop a whole organ shape. Here we report the dynamic, autonomous formation of the optic cup (retinal primordium) structure from a three-dimensional culture of mouse embryonic stem cell aggregates. Embryonic-stem-cell-derived retinal epithelium spontaneously formed hemispherical epithelial vesicles that became patterned along their proximal–distal axis. Whereas the proximal portion differentiated into mechanically rigid pigment epithelium, the flexible distal portion progressively folded inward to form a shape reminiscent of the embryonic optic cup, exhibited interkinetic nuclear migration and generated stratified neural retinal tissue, as seen *in vivo*. We demonstrate that optic-cup morphogenesis in this simple cell culture depends on an intrinsic self-organizing program involving stepwise and domain-specific regulation of local epithelial properties.

Eye formation has attracted the attention of both classical and modern-day developmental biologists[1–12]. The retinal anlage, which is demarcated by Rx (also called Rax) expression[13,14], first appears as the optic vesicle, an epithelial vesicle evaginating laterally from the diencephalon. Subsequently, its distal portion invaginates to form a two-walled cup-like structure, the optic cup (Supplementary Fig. 1a), which develops into the outer (pigmented) and inner (neurosensory) layers of the retina. Optic-cup development occurs in a complex environment affected by many neighbouring tissues. Therefore, its mechanistic analysis is far from simple, and such complexities have led to controversial results[9,10], in particular, with regard to the requirement of the surface ectoderm and lens in the invagination of the neural retina[15–18].

Here, we took an experimental approach using a three-dimensional (3D) embryonic stem (ES) cell culture system[19,20] and successfully reduced the complexity of this organogenetic process.

Optic-cup self-formation in 3D ES cell culture

We previously demonstrated the self-formation of stratified cerebral cortical tissues in culture, where floating aggregates of ES cells were cultured under low growth-factor conditions (serum-free floating culture of embryoid-body-like aggregates with quick reaggregation or SFEBq)[20,21]. We could also induce retinal differentiation in a modified SFEBq culture using transient activin treatment[22], but no clear formation of retinal epithelial structures was observed (data not shown). We then further modified the culture medium (see Methods) and added basement-membrane matrix components (matrigel) to promote the formation of rigid continuous epithelial structure[23] (Supplementary Fig. 1b–d). These modifications successfully improved the efficiency of retinal induction, resulting in 30–70% Rx–GFP$^+$ cells in total cells (typically, ~80% of aggregates were strongly positive for Rx–GFP; see Supplementary Fig. 1e for the dependence on integrin signals). A similar high percentage of Rx induction was induced by treating ES cell aggregates with defined matrix proteins (purified laminin and entactin) in the presence of the activin-family protein Nodal during days 1–7 (Supplementary Fig. 1f).

By day 6, initially homogenous ES cell aggregates formed hollowed spheres consisting of polarized N-cadherin$^+$ neuroepithelium with the apical surface inside (Supplementary Fig. 1g–j), which spontaneously subdivided into Rx–GFP$^+$ and Rx–GFP$^-$ portions (Fig. 1a and Supplementary Movie 1). On day 7, the Rx–GFP$^+$ portions formed hemispherical epithelial vesicles evaginating from the main body with one to four vesicles per aggregate (73.7 ± 2.5% Rx$^+$ spheres; Fig. 1b, c). The formed Rx–GFP$^+$ tissues were Pax6$^+$Sox1$^-$ (Fig. 1d, e), consistent with retinal marker expression[19,22], whereas the Rx–GFP$^-$ portions were the non-retinal neuroectodermal epithelium (Sox1$^+$; Fig. 1e).

Notably, on days 8–10, the Rx–GFP$^+$ vesicles underwent a dynamic shape change and formed a two-walled cup-like morphology (57.0 ± 4.7% of vesicles, Fig. 1f, g and Supplementary Fig. 1k, l; typically 200–400 μm in diameter; the optic cup in the E10.5 mouse is ~300 μm). The distal portion of the vesicle invaginated and expressed neural retina markers[13,14,24] such as Chx10 (also called Vsx2) and Six3 in addition to Rx and Pax6 (Fig. 1h, i and not shown). In contrast, the proximal epithelium, now forming the outer shell, expressed Pax6, Mitf, Coup-TF2 (also called Nr2f2), Otx2 and connexin 43, and a low level of Rx (Fig. 1i, j and Supplementary Fig. 1m–p), reminiscent of the marker profile of retinal pigment epithelium (RPE) progenitors[4,25]. This outer portion subsequently became pigmented (Fig. 1k). Like the one *in vivo*, the ES-cell-derived optic cup had a clear apical (aPKC)–basal (laiminin) polarity (Fig. 1l), and the inner portion invaginated with its apical side convex.

Importantly, the formation of the optic cup in the ES cell culture occurred in the absence of a lens (no crystalline$^+$ tissue; $n = 50$ aggregates) or surface ectodermal tissues (see the *in vivo* situation in Fig. 1m and Supplementary Fig. 1q; compare to Fig. 1h), indicating that the invagination was not caused by external structures but occurred in a self-directed fashion (Fig. 1n). Hydrostatic pressure seems to have no

[1] Organogenesis and Neurogenesis Group, RIKEN Center for Developmental Biology, Kobe 650-0047, Japan. [2] Four-Dimensional Tissue Analysis Unit, RIKEN Center for Developmental Biology, Kobe 650-0047, Japan. [3] Department of Biomechanics, Institute for Frontier Medical Sciences, Kyoto University, Kyoto 606-8507, Japan. [4] Laboratory of Extracellular Matrix Biochemistry, Institute for Protein Research, Osaka University, Suita 565-0871, Japan. [5] Computational Cell Biomechanics Team, VCAD System Research Program, RIKEN, Wako 351-0198, Japan.

©2011 Macmillan Publishers Limited. All rights reserved

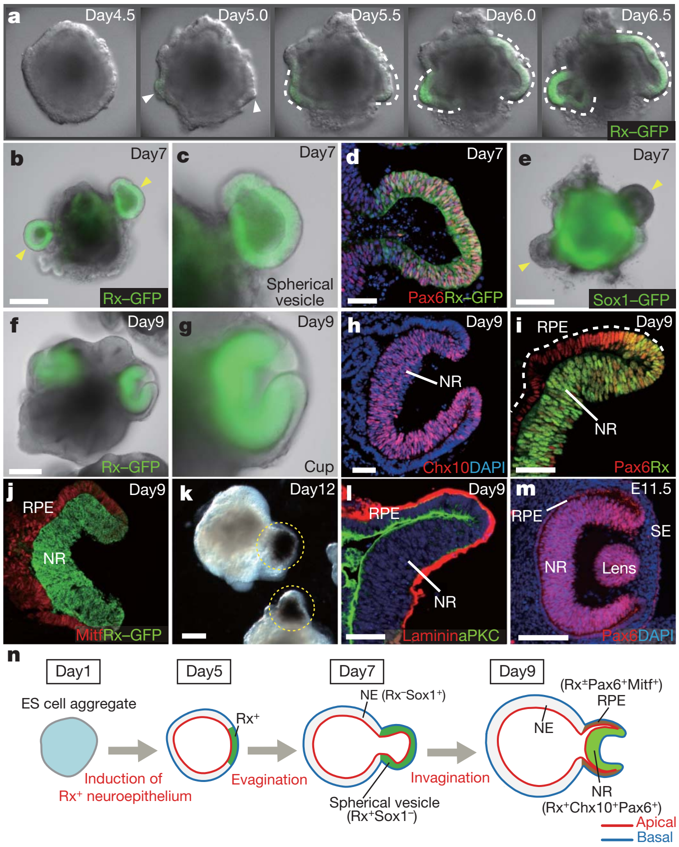


Figure 1 | Self-formation of an optic-cup-like structure in 3D culture of ES cell aggregates. **a**–**e**, SFEBq/matrigel culture. **a**–**e**, Self-formation of vesicles expressing Rx–GFP (**a**–**c**) and Pax6 (**d**), but not Sox1–GFP (**e**). **f**–**i**, Rx–GFP$^+$ eye-cup structures on day 9 (strongly and weakly in the neural retina (NR) and RPE portions, respectively). **h**, **i**, The inner portion strongly expressed Chx10 (**h**) and Rx and moderately expressed Pax6 (**i**) on day 9. **j**, **k**, The outer epithelial shell of the cup expressed Mitf (**j**; day 9) and accumulated pigment on day 12 (**k**). **l**, The apical marker aPKC and laminin$^+$ basement membrane. **m**, E11.5 mouse eye. **n**, Schematic of optic-cup self-formation. SE, surface ectoderm. Scale bars: 200 μm (**b**, **e**–**f**, **k**), 50 μm (**d**, **h**, **i**, **l**), 100 μm (**m**).

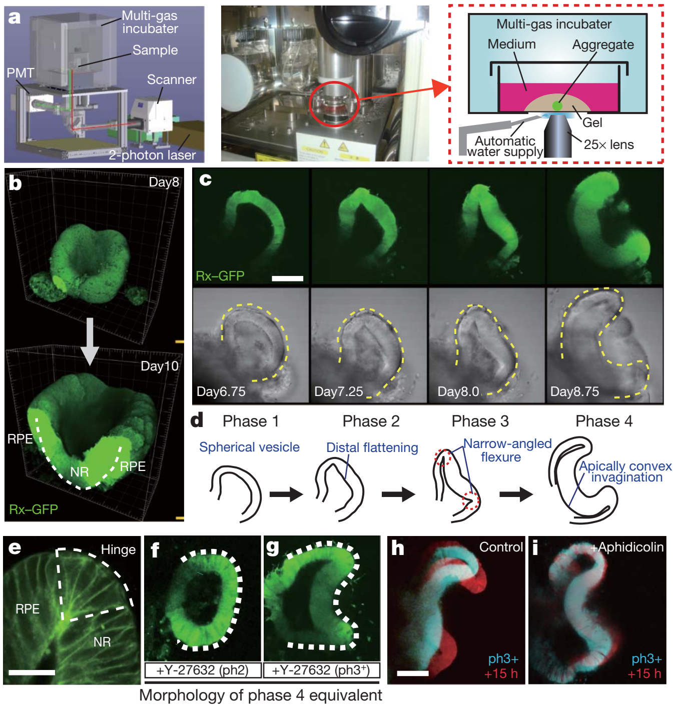


Figure 2 | Progressive morphogenetic changes of ES-cell-derived retinal epithelium. **a**, Multi-photon device for long-term 3D live imaging. PMT, photomultiplier. **b**, Surface-rendering 3D reconstruction images of invagination. The front part (bright green) is a cross-section image. **c**, Optical cross-section of Rx–GFP images (top) and laser-scanned bright-field images (bottom; dotted lines indicate the basal side). **d**, Four phases during the invagination process. **e**, Multi-photon optical section near the hinge at phase 4. **f**, **g**, Effects of Y-27632 treatment (**f**, starting from phase 2 (ph2); **g**, from late phase 3 (ph3$^+$)) on invagination. **h**, **i**, Treatment of the phase 3$^+$ neural retina (blue) with (**i**) or without (**j**) aphidicolin. Red, 15 h later. Scale bars: 100 μm (**c**, **h**), 20 μm (**e**).

essential role, as opening a small hole through the neuroectodermal epithelium did not substantially affect eye-cup formation.

Four phases of 3D eye-cup morphogenesis

We then continued to analyse the morphogenetic process in 3D, using a specially assembled multi-photon live-imaging system in combination with a full-sized CO_2/O_2 incubator (Fig. 2a and Supplementary Fig. 2a, b), which allows a constant, healthy culture environment for a week or longer (see Fig. 2b and Supplementary Movie 2, part a, for an example of the time-lapse deep imaging). The invagination process consisted of four consecutive phases (Fig. 2c, d and Supplementary Movie 2, parts b, c), consistent with *in vivo* development[26]. On day 6, the evaginated vesicle was hemispherical in shape (phase 1). In phase 2, occurring around day 7, the distal portion of the vesicle became flattened (Supplementary Fig. 2c). Subsequently, in phase 3, the angle at the joint (hinge) between the neural retina and RPE domains became narrower or even acute (Fig. 2e). Then, on day 8, the neural retina epithelium started to expand as an apically convex structure, forming a cup via progressive invagination (phase 4).

The microfilament system has been implicated as a key internal regulator in various aspects of epithelial morphogenesis and is often regulated by the Rho-ROCK system[27]. The invagination morphogenesis (including flattening and hinge formation) was blocked when the culture was treated with the ROCK inhibitor Y-27632 (ref. 28) (which is known to attenuate myosin activity) from phase 1 or 2 (Fig. 2f and data not shown; invagination 0%, $n = 30$, both phases). A similar effect was seen with the myosin inhibitor blebbistatin (data not sown). In contrast, when Y-27632 was added after the hinge appeared, the neural retina reproducibly continued to invaginate at phase 4 ($40 \pm 3.9\%$, $n = 30$, three experiments; Fig. 2g), indicating a phase-specific dependence (phases 1–3) of morphogenesis on ROCK activity.

At phase 4, the invaginating neural retina epithelium expanded substantially both in the tangential and vertical directions, whereas the RPE domain (particularly in its distal part) expanded exclusively tangentially (Supplementary Fig. 2e–g). Treatment with the mitotic inhibitor aphidicolin[29] from late phase 3 onwards efficiently blocked the invagination progress during phase 4 (Fig. 2h, i and Supplementary Movie 3) as well as tissue expansion and cell proliferation (Supplementary Fig. 2h–j), indicating that tissue expansion has an essential role in the phase-4 morphogenesis.

Stepwise acquisition of domain-specific properties

Detailed 3D analysis showed that, whereas the entire phase-1 vesicle consisted of simple columnar epithelium, the epithelial cells of the RPE, neural retina and hinge domains at phase 4 exhibited distinct morphologies, reminiscent of their shapes *in vivo*[26]: columnar epithelial cells, pseudo-stratified cells, and apically narrow wedge-shaped cells, respectively (Fig. 3a, Supplementary Fig. 2d, bottom, and Supplementary Movie 2, part d; no wedge-shaped cells appeared at the hinge in culture pre-treated with either Y-27632 or blebbistatin; data not shown).

Both *in vitro* and *in vivo*, neural retina invagination occurs via unique apically convex invagination, in contrast to the well studied apically concave invagination seen in vertebrate neural-tube formation

©2011 Macmillan Publishers Limited. All rights reserved

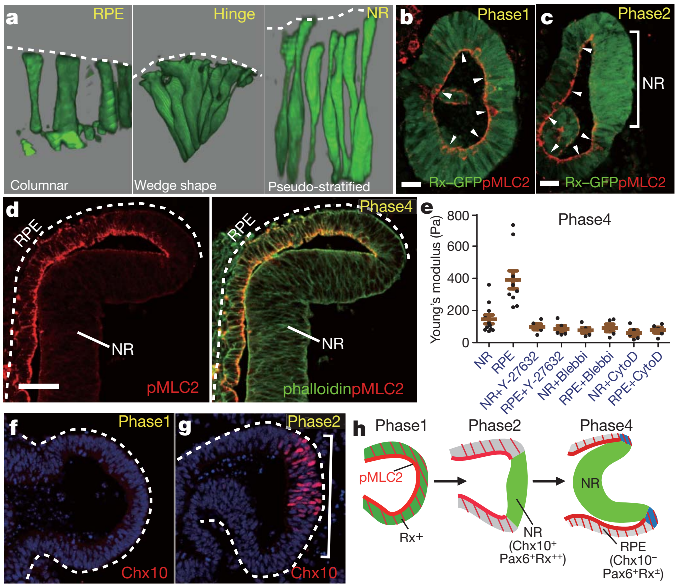


Figure 3 | Stepwise acquisition of domain-specific epithelial properties. **a**, Three-dimensional reconstruction views of the RPE, hinge and neural retina cells at phase 4. **b**, High pMLC2 levels (arrowheads) were seen in the entire Rx–GFP$^+$ epithelium at phase 1. **c**, Reduced pMLC2 accumulation in the neural retina (bracket) at phase 2. **d**, Phase 4. Reduced and high pMLC levels in the neural retina and in the RPE and hinge regions, respectively. **e**. AFM-based measurement of relative tissue rigidity of phase 4 neural retina and RPE tissues with or without Y-27632, blebbistatin and cytochalasin D treatments. **f**, **g**, Expression of Chx10 at phase 1 (**f**) and phase 2 (**g**). **h**, Schematic summary. Scale bars: 50 μm (**b**–**d**). Error bars in **e** represent s.e.m.

and during fly gastrulation, where ROCK-myosin activation is essential[27,28,30–34]. With this in mind, we next examined levels of phosphorylated myosin light chain 2 (pMLC2; MLC2 is also called Myl2), which reflects local actomyosin activation[30], in different domains and at different phases. The proximal and distal epithelia of the phase-1 vesicle showed comparable levels of pMLC2 (Fig. 3b and Supplementary Fig. 3a). This pMLC2 accumulation seemed to depend on the Rho-ROCK system, as Y-27632 diminished pMLC2 levels (data not shown; consistent with this, the inner space of the phase-1 vesicle gradually decreased whereas Y-27632 reversed this reduction and expanded the whole vesicle evenly; Supplementary Fig. 3b). In contrast to the even accumulation at phase 1, the pMLC2 levels at phases 2–3 were reduced in the neural retina but remained high in the RPE and hinge epithelium (Fig. 3c and Supplementary Fig. 3c). The pMLC2 levels of the neural retina became particularly low at phase 4 (Fig. 3d and Supplementary Fig. 3d). The reduced pMLC2 level was also observed in the neural retina of the mouse embryo (Supplementary Fig. 3e, f). Consistent with the reduced level of pMLC, the mechanical rigidity of the RPE tissue in the phase-4 optic cup was substantially higher than that of the neural retina as measured by a force probe using an atomic force microscope (AFM) cantilever[35,36], whereas this difference disappeared after treatment with ROCK or actomyosin inhibitors (Fig. 3e and Supplementary Fig. 3g). In contrast, a substantial level of actomyosin-dependent tissue rigidity was seen in both the distal and proximal epithelia during phase 1 (Supplementary Fig. 3h and not shown).

We also observed differential domain- and phase-specific mechanical properties in response to cell ablation. When a few cells were ablated[37] by means of a 3D-targeted multi-photon laser beam (Supplementary Fig. 3i–l), the ablated space in the phase-4 RPE and neural retina, but not in the hinge, soon disappeared by compression from neighbouring cells within the epithelium (Supplementary Movie 4, parts a–c). Mitotic inhibitor treatment inhibited the gap-filling behaviour (Supplementary Movie 4, part d), indicating an essential role for cell proliferation in generating the pushing force. In contrast, at phases 1–2, the ablation-induced gaps, whether in the distal or proximal epithelia, did not close after ablation but began to increase, presumably by local pulling tension (part e and not shown), as in the hinge during phase 4.

As for marker expression, Chx10 expression started by phase 2 in the distal retinal epithelium, whereas the proximal portion was Chx10$^-$Pax6$^+$ and showed decreased Rx expression (Fig. 3f–h; see *in vivo* expression in Supplementary Fig. 3m). Collectively, these findings demonstrate that the self-formed retinal epithelium sequentially acquires quantitatively distinct morphological, biochemical, mechanical and gene-expression properties in a domain-specific manner during early eye-cup morphogenesis.

Self-patterning into neural retina and RPE domains

We next asked whether the neural retina and RPE fates were irreversibly determined when vesicle evagination was complete. To test this, we excised Rx$^+$ phase-1 vesicles with or without neighbouring neuroectodermal epithelium tissue (Rx$^-$) from the day-7 ES cell aggregates (Fig. 4a–k). A few days later, the Rx$^+$ vesicles excised with a small amount of non-retinal neuroectodermal epithelium developed into spatially ordered RPE and neural retina epithelium, and even formed an optic-cup structure (Fig. 4a–c, j, k). In contrast, Rx$^+$ epithelium isolated without non-retinal tissues failed to invaginate, and predominantly expressed Chx10 on day 10 (Fig. 4d–f; bracket), whereas only a small number of cells were Chx10$^-$Pax6$^+$ (arrowheads).

Canonical Wnt signalling has been implicated in the specification and maintenance of RPE fate[38–41]. When phase-1 vesicles isolated without neuroectodermal epithelium tissue were co-cultured with an aggregate of Wnt3a-expressing L cells (Fig. 4g and Supplementary Fig. 4a, b), both the RPE differentiation and neural retina invagination were substantially rescued (Fig. 4h–k; Supplementary Movie 5, using a disc-confocal incubation microscope in Supplementary Fig. 2k), whereas such rescue was not observed with control L cells (not shown). A similar rescue was observed in the isolated vesicle culture to which Wnt3a protein was added (Fig. 4j, k and Supplementary Fig. 4c, d). Accordingly, treating the whole aggregates with the Wnt secretion inhibitor IWP2 interfered with invagination, reduced RPE tissues and increased the percentage of Chx10$^+$ tissues (Fig. 4l, m and Supplementary Fig. 4e–h; the effects of IWP2 on differentiation were reversed by adding Wnt3a, whereas no substantial effects were seen on proliferation or apoptosis; Supplementary Fig. 4i, j). Conversely, the Wnt agonist BIO (6-bromoindirubin-3′oxime; GSK3-β inhibitor) strongly promoted RPE differentiation and suppressed neural retina generation (Fig. 4l, m and Supplementary Fig. 4k, l).

Taken together, these findings indicate that RPE differentiation requires tissue interactions involving diffusible induction factors[42] from neuroectodermal epithelium. In contrast, the neural retina fate is tissue-autonomously determined; this might, in part, result from Rx and Six3, which attenuate Wnt pathway activation[40], and also from Wnt antagonists such as Dkk1 (ref. 43), which is strongly expressed in the retinal tissues (Supplementary Fig. 4m–q). Interestingly, even at phase 1, *Dkk1* expression in the vesicle already showed a biased distally high expression (Supplementary Fig. 4r), which may partly explain why the addition of soluble Wnt3a protein in the explant culture could also spatially rescue the RPE–neural retina pattern (Supplementary Fig. 4d).

Self-directed stratification of neural retina tissue

In the growing phase-4 neural retina, the progenitors underwent frequent cell divisions coupled with interkinetic nuclear migration (Fig. 5a, Supplementary Fig. 5a and Supplementary Movie 6, parts a, b), as seen *in vivo*[44]. This led us to examine the formation of tissue architecture in isolated neural retina epithelia in long-term culture. When phase-4 eye cups were manually excised en bloc at the RPE domain from the aggregates, the RPE portions flipped over like a leaf

©2011 Macmillan Publishers Limited. All rights reserved

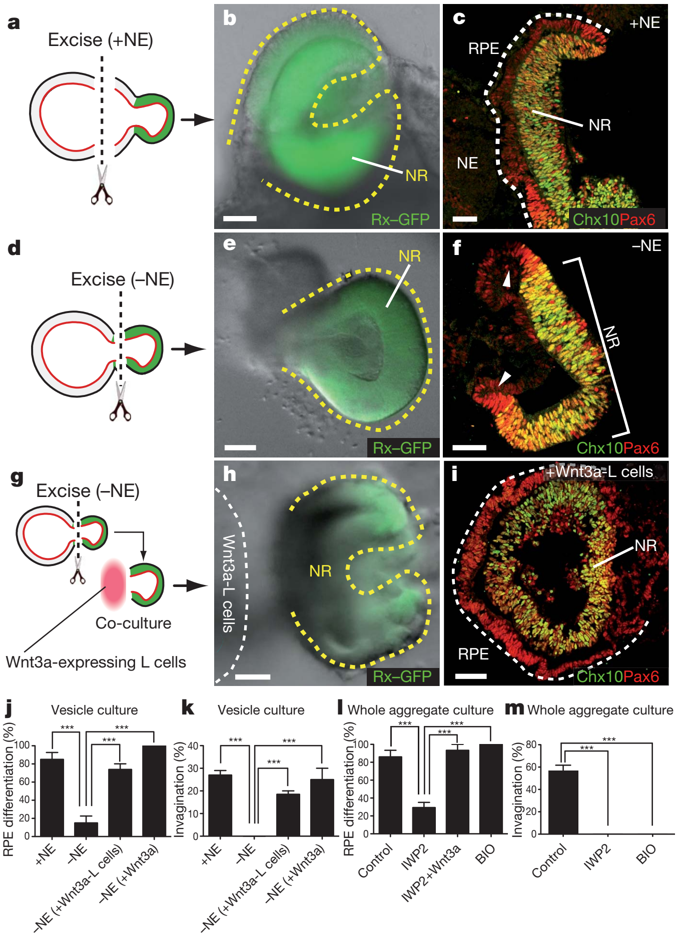


Figure 4 | Self-patterning into neural retina and RPE via interactions with neuroectodermal epithelium. a–k, Excision studies. The phase 1 optic vesicles were isolated manually with (**a–c**) or without (**d–f**) non-retinal neuroectodermal epithelium and cultured for 3 days. **g–i**, Isolated vesicles were co-cultured with Wnt3a-expressing L cells re-aggregated in a drop of high-concentration matrigel. **b**, **e**, **h**, Differential interference contrast images with Rx–GFP fluorescence. **c**, **f**, **i**, Immunostaining with Chx10 and Pax6 (RPE tissues, Pax6$^+$Chx10$^-$). **j**, **k**, Quantification of balanced neural retina versus RPE differentiation (**j**; RPE/neural retina+RPE>0.5) and the frequencies of successful neural retina invagination (**k**). **l**, **m**, Effects of the Wnt inhibitor IWP2 and the agonist BIO on the neural retina versus RPE specification (**l**) and neural retina invagination (**m**) in non-excised SFEBq aggregates. Error bars represent s.e.m. (three independent experiments using 30 aggregates each in **j**, **k**.). ***, $P < 0.001$ (Tukey's test). Scale bars: 50 μm (**b**, **c**, **e**, **f**, **h**, **i**).

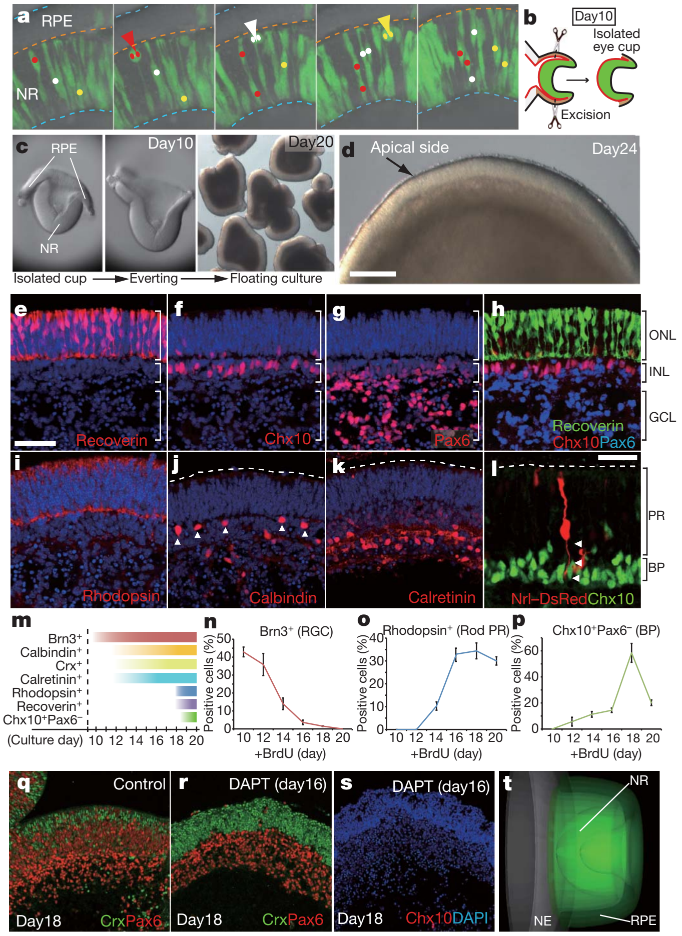


Figure 5 | Generation of stratified neural retina tissues from ES-cell-derived invaginated epithelia. a, Cell tracing of phase-4 neural retina progenitors is shown (the coloured dots indicate nuclear position; ~11 h). **b**, **c**, Isolation of the phase-4 optic cups from the SFEBq aggregates at day 10. Eversion of the RPE hinge occurring after excision (**c**; middle). Floating tissues on day 20 are shown (**c**; right). **d**, Uniformly extending retinal epithelium in day 24 culture. **e–k**, Neural retina markers in cross-sections on day 24. BP, bipolar cells; GCL, ganglion cell layer; INL, inner nuclear layer; ONL, outer nuclear layer; PR, photoreceptors. **l**, pNrl-DsRed2 was electroporated (apically) on day 16, and analysed on day 24. **m**, Temporal expression profile of neural retina markers. **n–p**, BrdU was incorporated on the given day and the BrdU-retaining cell types were analysed on day 24. **q–t**, Expression of Crx and Pax6 (**q**, **r**) and Chx10 (**s**) in day-18 culture with (**r**, **s**) or without (**q**) DAPT treatment. **t**, Three-dimensional perspective of computer-simulated phase-4 eye cup based on our working model described in 'Discussion'. Scale bars: 100 μm (**d**), 50 μm (**e**), 20 μm (**l**). Error bars represent s.e.m.

spring at the hinge and everted in an inside-out fashion (Fig. 5b, c, Supplementary Fig. 5a and Supplementary Movie 7; such eversion was not seen in the samples shown in Fig. 4g presumably because their cut ends spontaneously became sealed soon after). In contrast to RPE, the shape of neural retina did not substantially change, demonstrating the tissue's plasticity; in fact, this mechanical property was useful for subjecting the isolated phase-4 neural retina epithelia to further suspension culture in retina maturation medium[45] (Supplementary Fig. 5b).

Notably, 10–14 days later, these isolated tissues spontaneously formed large, continuous epithelial structures with clear stratification (Fig. 5d; typically 800–2,000 μm in diameter), reminiscent of the early postnatal retina consisting of multiple layers of distinct cellular components[5,13] (Supplementary Fig. 5c). The ES-cell-derived stratified retinal tissues contained all of the major neural retina components (Fig. 5e–k), that is, photoreceptors (rhodopsin$^+$, recoverin$^+$; Fig. 5e, h, i and Supplementary Fig. 5d), ganglion cells (Brn3$^+$Pax6$^+$calretinin$^+$; Fig. 5g, h, k and Supplementary Fig. 5e), bipolar cells (Chx10$^+$Pax6$^-$, Fig. 5f–h; Chx10 expression is bipolar-cell-specific at this late stage), horizontal cells (calbindin$^+$calretinin$^-$; Fig. 5j), amacrine cells (Pax6$^+$ calretinin$^+$; Fig. 5g, k) and Muller glia (CRALBP$^+$; Supplementary Fig. 5f). These neural retina components are spatially arranged into the correct apical–basal order as seen in the early neonatal eye. Photoreceptors comprised the outermost layer, overlaying a layer of bipolar cells (Fig. 5h, l). In Fig. 5l, the shape of a photoreceptor cell is visualized with DsRed driven by the photoreceptor-specific *Nrl* promoter[46] (Supplementary Fig. 5g–i: *Nrl* expression started just before the terminal cell division of photoreceptors and became strong soon after; Supplementary Movie 6, part c). Horizontal cells were found between the photoreceptor and bipolar layers. The innermost zone contained ganglion cells and amacrine cells, and the formation of an inner-plexiform-layer-like zone was seen there[47] (Supplementary Fig. 5j, arrow, and k).

As a minor difference, the ganglion cells in this culture tended to scatter rather than aligning as *in vivo*, probably because there was no

©2011 Macmillan Publishers Limited. All rights reserved

rigid border structure such as that provided by the inner limiting membrane *in vivo*. In addition, unlike rods, cone photoreceptors (expressing colour opsins) were present in a relatively low percentage ($<$0.1%; Supplementary Fig. 5l, m and data not shown).

The expression profile of stage-specific differentiation markers of neural retina components, summarized in Fig. 5m, indicated that the temporal order in ES-cell culture followed that *in vivo*. In generation-date analysis[20], the cell-cycle exit of ganglion cells (Brn3$^+$), photoreceptors (rhodopsin$^+$) and bipolar cells (Chx10$^+$Pax6$^-$) peaked around days 10, 16 and 18, respectively, which is in accordance with the *in vivo* generation order[13] (Fig. 5n–p and Supplementary Fig. 5n, o). Consistent with these observations, treatment with the Notch inhibitor DAPT (on day 16), which promotes on-going differentiation and inhibits proliferation, increased the number of Crx$^+$ photoreceptors on day 18 whereas it markedly reduced the number of Chx10$^+$ cells (mitotic neural retina progenitors and postmitotic bipolar cells) and Ptf1a$^+$ GABAergic precursors (Fig. 5q–s and Supplementary Fig. 5p–s, and data not shown).

Collectively, these findings demonstrate that the fully stratified neural retina tissue architecture in this ES-cell culture self-forms in a spatiotemporally regulated manner mimicking *in vivo* development.

Discussion

The self-directed organization of the complex optic-cup and neural retina morphology, shown here, was unanticipated, as the ES-cell culture started as patternless aggregates of homogeneous pluripotent cells, and continued under a uniform culture environment. Thus, this system offers a representation of the emergence of collective cell behaviour in a non-equilibrium biological process that evolves spatiotemporally. The invaginaiton of the neural retina could occur even without forces from external structures at least in this *in vitro* context. Whereas the complete elucidation of the mechanics of this autonomous invagination requires further extensive investigation, this study indicates that the morphogenesis includes multiple steps controlled in a history-dependent fashion.

Following the formation of the hinge, phase-3 neural retina tends to have a slight apical bending. A working model for future investigation, consistent with our findings for tissue movement and mechanical properties during phase 4, is that, surrounded by the less-yielding RPE shell (pMLC2-high) and in a limited space, the flexible phase-4 neural retina is promoted to bend further (buckle apically) by its tangential expansion. Indeed, such a morphogenetic process could also be recapitulated by computer simulation based on the following three sequential local rules consistent with our observations: (1) relaxation of the presumptive neural retina beginning at phase 2; (2) apical constriction in the hinge epithelium at phase 3; and (3) rapid tangential growth of the neural retina and the distal RPE tissues at phase 4. In this model involving the local control of apical, basal and transmural springs, the neural retina spontaneously involutes inside the RPE shell, following flattening and hinge formation, conceptually supporting the feasibility of spontaneous cup morphogenesis independent of forces from external structures (Fig. 5t, Supplementary Fig. 6a, b and Supplementary Movie 8; see also Supplementary Fig. 6c–e for response to perturbations).

This study has revealed that the complex morphogenesis of the retinal anlage, at least in the *in vitro* context, possesses a 'latent intrinsic order' involving dynamic self-patterning and self-formation driven by a sequential combination of local rules and internal forces within the epithelium. The *in vivo* situation should be certainly more complex, and extrinsic signals and forces from external structures (for example, the surface ectoderm, lens and periocular mesenchyme)[9,10,42,48,49] as well as space constraints presumably work together with this intrinsic order to reinforce robust retinal morphogenesis.

With regard to the application aspect, self-formation of fully stratified 3D neural retina tissues heralds the next-generation of generative medicine in retinal degeneration therapeutics, and opens up new avenues for the transplantation of artificial retinal tissue sheets[50], rather than simple cell grafting.

METHODS SUMMARY

The SFEBq culture was performed as described previously[20], and matrigel (final 2%) was added to the 1.5% KSR-containing differentiation medium on day 1. Multi-photon imaging and ablation were performed using 900-nm femtosecond laser pulses (80 MHz) from a MaiTai eHP (Spectra-Physics) using a ×25 water-immersion objective lens (NA 1.05; Olympus).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 30 November 2010; accepted 17 February 2011.

1. Spemann, H. Ueber korrelationen in der entwicklung des auges. *Verh. Anat. Ges.* **15,** 61–79 (1901).
2. Lewis, W. H. Experimental studies on the development of the eye in Amphibia I. On the origin of the lens in *Rana palustrius*. *Am. J. Anat.* **3,** 505–536 (1904).
3. Nakagawa, S., Takada, S., Takada, R. & Takeichi, M. Identification of the laminar-inducing factor: Wnt-signal from the anterior rim induces correct laminar formation of the neural retina *in vitro*. *Dev. Biol.* **260,** 414–425 (2003).
4. Martinez-Morales, J. R., Rodrigo, I. & Bovolenta, P. Eye development: a view from the retina pigmented epithelium. *Bioessays* **26,** 766–777 (2004).
5. Mu, X. & Klein, W. H. A gene regulatory hierarchy for retinal ganglion cell specification and differentiation. *Semin. Cell Dev. Biol.* **15,** 115–123 (2004).
6. Wilson, S. W. & Houart, C. Early steps in the development of the forebrain. *Dev. Cell* **6,** 167–181 (2004).
7. Rembold, M., Loosli, F., Adams, R. J. & Wittbrodt, J. Individual cell migration serves as the driving force for optic vesicle evagination. *Science* **313,** 1130–1134 (2006).
8. Cayouette, M., Poggi, L. & Harris, W. A. Lineage in the vertebrate retina. *Trends Neurosci.* **29,** 563–570 (2006).
9. Adler, R. & Canto-Soler, M. V. Molecular mechanisms of optic vesicle development: complexities, ambiguities and controversies. *Dev. Biol.* **305,** 1–13 (2007).
10. Martinez-Morales, J. R. & Wittbrodt, J. Shaping the vertebrate eye. *Curr. Opin. Genet. Dev.* **19,** 511–517 (2009).
11. Picker, A. *et al.* Dynamic coupling of pattern formation and morphogenesis in the developing vertebrate retina. *PLoS Biol.* **7,** e1000214 (2009).
12. Fuhrmann, S. Eye morphogenesis and patterning of the optic vesicle. *Curr. Top. Dev. Biol.* **93,** 61–84 (2010).
13. Livesey, F. J. & Cepko, C. L. Vertebrate neural cell-fate determination: lessons from the retina. *Nature Rev. Neurosci.* **2,** 109–118 (2001).
14. Bailey, T. J. *et al.* Regulation of vertebrate eye development by Rx genes. *Int. J. Dev. Biol.* **48,** 761–770 (2004).
15. Lopashov, G. *Developmental Mechanism of Vertebrate Eye Rudiments* (Pergammon, 1963).
16. Hyer, J., Mima, T. & Mikawa, T. FGF1 patterns the optic vesicle by directing the placement of the neural retina domain. *Development* **125,** 869–877 (1998).
17. Hyer, J., Kuhlman, J., Afif, E. & Mikawa, T. Optic cup morphogenesis requires pre-lens ectoderm but not lens differentiation. *Dev. Biol.* **259,** 351–363 (2003).
18. Smith, A. N., Miller, L. A., Radice, G., Ashery-Padan, R. & Lang, R. A. Stage-dependent modes of Pax6-Sox2 epistasis regulate lens development and eye morphogenesis. *Development* **136,** 2977–2985 (2009).
19. Wataya, T. *et al.* Minimization of exogenous signals in ES cell culture induces rostral hypothalamic differentiation. *Proc. Natl Acad. Sci. USA* **105,** 11796–11801 (2008).
20. Eiraku, M. *et al.* Self-organized formation of polarized cortical tissues from ESCs and its active manipulation by extrinsic signals. *Cell Stem Cell* **3,** 519–532 (2008).
21. Au, E. & Fishell, G. Cortex shatters the glass ceiling. *Cell Stem Cell* **3,** 472–474 (2008).
22. Ikeda, H. *et al.* Generation of Rx$^+$/Pax6$^+$ neural retinal precursors from embryonic stem cells. *Proc. Natl Acad. Sci. USA* **102,** 11331–11336 (2005).
23. Fujiwara, H. *et al.* Regulation of mesodermal differentiation of mouse embryonic stem cells by basement membranes. *J. Biol. Chem.* **282,** 29701–29711 (2007).
24. Lagutin, O. *et al.* Six3 promotes the formation of ectopic optic vesicle-like structures in mouse embryos. *Dev. Dyn.* **221,** 342–349 (2001).
25. Tang, K. *et al.* COUP-TFs regulate eye development by controlling factors essential for optic vesicle morphogenesis. *Development* **137,** 725–734 (2010).
26. Hilfer, S. R. & Yang, J.-J. W. Accumulation of CPC-precipitable material at apical cell surfaces during formation of the optic cup. *Anat. Rec.* **197,** 423–433 (1980).
27. Sawyer, J. M. *et al.* Apical constriction: a cell shape change that can drive morphogenesis. *Dev. Biol.* **341,** 5–19 (2010).
28. Kinoshita, N., Sasai, N., Misaki, K. & Yonemura, S. Apical accumulation of Rho in the neural plate is important for neural plate cell shape change and neural tube formation. *Mol. Biol. Cell* **19,** 2289–2299 (2008).
29. Agius, E. *et al.* Converse control of oligodendrocyte and astrocyte lineage development by Sonic hedgehog in the chick spinal cord. *Dev. Biol.* **270,** 308–321 (2004).
30. Amano, M. *et al.* Phosphorylation and activation of myosin by Rho-associated kinase (Rho-kinase). *J. Biol. Chem.* **271,** 20246–20249 (1996).
31. Schoenwolf, G. C. & Smith, J. L. Mechanisms of neurulation: traditional viewpoint and recent advances. *Development* **109,** 243–270 (1990).
32. Haigo, S. L., Hildebrand, J. D., Harland, R. M. & Wallingford, J. B. Shroom induces apical constriction and is required for hingepoint formation during neural tube closure. *Curr. Biol.* **13,** 2125–2137 (2003).

©2011 Macmillan Publishers Limited. All rights reserved

33. Hildebrand, J. D. Shroom regulates epithelial cell shape via the apical positioning of an actomyosin network. *J. Cell Sci.* **118,** 5191–5203 (2005).
34. Nishimura, T. & Takeichi, M. Shroom3-mediated recruitment of Rho kinases to the apical cell junctions regulates epithelial and neuroepithelial planar remodeling. *Development* **135,** 1493–1502 (2008).
35. Rico, F. *et al.* Probing mechanical properties of living cells by atomic force microscopy with blunted pyramidal cantilever tips. *Phys. Rev. E* **72,** 021914 (2005).
36. Krieg, M. *et al.* Tensile forces govern germ-layer organization in zebrafish. *Nature Cell Biol.* **10,** 429–436 (2008).
37. Mascaro, A. L., Sacconi, L. & Pavone, F. S. Multi-photon nanosurgery in live brain. *Front. Neuroenergetics* **2,** (2010).
38. Fuhrmann, S. Wnt signaling in eye organogenesis. *Organogenesis* **4,** 60–67 (2008).
39. Westenskow, P., Piccolo, S. & Fuhrmann, S. β-catenin controls differentiation of the retinal pigment epithelium in the mouse optic cup by regulating Mitf and Otx2 expression. *Development* **136,** 2505–2510 (2009).
40. Liu, W., Lagutin, O., Swindell, E., Jamrich, M. & Oliver, G. Neuroretina specification in mouse embryos requires Six3-mediated suppression of Wnt8b in the anterior neural plate. *J. Clin. Invest.* **120,** 3568–3577 (2010).
41. Fujimura, N., Taketo, M. M., Mori, M., Korinek, V. & Kozmik, Z. Spatial and temporal regulation of Wnt/β-catenin signaling is essential for development of the retinal pigment epithelium. *Dev. Biol.* **334,** 31–45 (2009).
42. Yang, X. J. Roles of cell-extrinsic growth factors in vertebrate eye pattern formation and retinogenesis. *Semin. Cell Dev. Biol.* **15,** 91–103 (2004).
43. Diep, D. B., Hoen, N., Backman, M., Machon, O. & Krauss, S. Characterisation of the Wnt antagonists and their response to conditionally activated Wnt signalling in the developing mouse forebrain. *Brain Res. Dev. Brain Res.* **153,** 261–270 (2004).
44. Norden, C., Young, S., Link, B. A. & Harris, W. A. Actomyosin is the main driver of interkinetic nuclear migration in the retina. *Cell* **138,** 1195–1208 (2009).
45. Pinzón-Duarte, G., Kohler, K., Arango-Gonzalez, B. & Guenther, E. Cell differentiation, synaptogenesis, and influence of the retinal pigment epithelium in a rat neonatal organotypic retina culture. *Vision Res.* **40,** 3455–3465 (2000).
46. Matsuda, T. & Cepko, C. L. Controlled expression of transgenes introduced by *in vivo* electroporation. *Proc. Natl Acad. Sci. USA* **104,** 1027–1032 (2007).
47. Stella, S. L. Jr, Li, S., Sabatini, A., Vila, A. & Brecha, N. C. Comparison of the ontogeny of the vesicular glutamate transporter 3 (VGLUT3) with VGLUT1 and VGLUT2 in the rat retina. *Brain Res.* **1215,** 20–29 (2008).
48. Fuhrmann, S., Levine, E. M. & Reh, T. A. Extraocular mesenchyme patterns the optic vesicle during early eye development in the embryonic chick. *Development* **127,** 4599–4609 (2000).
49. Cavodeassi, F. *et al.* Early stages of zebrafish eye formation require the coordinated activity of Wnt11, Fz5, and the Wnt/β-catenin pathway. *Neuron* **47,** 43–56 (2005).
50. Seiler, M. J. *et al.* Visual restoration and transplant connectivity in degenerate rats implanted with retinal progenitor sheets. *Eur. J. Neurosci.* **31,** 508–520 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to S. Nakanishi, S. Yonemura, R. Ladher, K. Muguruma, H. Inomata and M. Ohgushi for comments and to members of the Sasai laboratory for discussion. We also thank Olympus, particularly Y. Saito, M. Suzuki and Y. Imai, for their help and discussion regarding the design, assembly and optimized utility of the incubator-combined multi-photon and confocal optic systems, and T. Sugitate and N. Saito at JPK Instruments for technical advice on the AFM assay. This work was supported by grants-in-aid from MEXT (Y.S., M.E., T.A.), the Knowledge Cluster Initiative at Kobe and the S-Innovation Project (Y.S., K.S.), and the Leading Project for Realization of Regenerative Medicine (Y.S).

Author Contributions M.E. and Y.S. designed the project and wrote the manuscript. M.E., N.T., M.K. and E.S performed experiments. K.S. provided critical technical information on matrix experiments. H.I., S.O. and T.A. performed computer simulation by discussing details with M.E. and Y.S.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.E. (eiraku@cdb.riken.jp) or Y.S. (yoshikisasai@cdb.riken.jp).

©2011 Macmillan Publishers Limited. All rights reserved

METHODS

ES cell culture. Mouse ES cells (EB5, *Rx-Venus*, *Sox1-GFP*) were maintained as described[20]. For the pigmentation assay (Fig. 1k), an ES cell line with B6/129 background was used. For SFEBq culture, ES cells were dissociated to single cells in 0.25% trypsin-EDTA and quickly re-aggregated in differentiation medium (3,000 cells per 100 μl per well) in 96-well low-cell-adhesion plates (Lipidure Coat, NOF). Differentiation medium was G-MEM supplemented with 1.5% knockout serum replacement (KSR; Invitrogen) +0.1 mM non-essential amino acids (Invitrogen) + 1 mM pyruvate + 1 mM 2-mercaptoethanol. As shown in Supplementary Fig. 1d, the addition of 1.5% KSR was much more efficient in Rx induction than that of 10% KSR, a condition that had been used previously[22]. Defining the day on which the SFEBq culture was started as day 0, matrigel (growth-factor-reduced; BD Biosciences) was added to culture to final 2% (v/v) on day 1. Alternatively, purified laminin and entactin proteins (200 μg ml^{-1}; BD Biosciences) and nodal protein (1 μg ml^{-1}) were added on day 1. The culture was transferred to bacterial-grade plastic dishes on day 7 and further cultured in suspension in DMEM/F12 medium supplemented with the N2 additive under 40%-O_2/5%-CO_2 conditions. The Rx induction occurred without substantial differentiation of mesodermal or endodermal tissues. For the sake of simplicity and cost, the matrigel treatment was used in the experiments unless otherwise mentioned.

Signalling molecules and antagonists. Growth factors were purchased from R&D Systems and used at the concentration indicated in the text and legends. The Notch inhibitor DAPT was applied to the isolated neural retina tissues from day 16. On day 18, the cell aggregates were fixed and analysed by immunohistochemistry. The integrin-blocking experiment (Supplementary Fig. 1) was performed by treating ES cells with the anti-β1-integrin antibody or control IgM (10 μg ml^{-1}; BD Pharmingen) from day 1, before the addition of matrigel to the culture.

IWP2 (4 μM) and BIO (500 nM) also inhibited the neural retina invagination in the whole-head tissue culture of E8.75 mouse embryos. Consistent with the data shown in Fig. 4, they increased and decreased the amount of neural retina tissues in the eyes, respectively. In this case, however, these inhibitors also blocked the invagination and differentiation of the lens, demonstrating the limitation of the embryonic tissue explant assays and the advantage of the ES-cell culture system in terms of the simplicity in investigating the self-organizing nature of neural retina invagination.

Live imaging of optic-cup formation. The 3D live imaging was performed using specially assembled inverted microscopes (confocal or multi-photon) combined with a full-sized CO_2/O_2 incubator. The position of the ES cell aggregate or isolated vesicle was fixed in a drop of undiluted matrigel, which was then immersed in culture medium (described above) on a 3.5-cm glass-bottom dish. For confocal analysis, optical section images were obtained using a ×20 objective lens (Olympus), a spinning disk confocal system (CSU-X1, Yokogawa) and an EM-CCD camera (Andor, 512 × 512 pixels; the incubation system is based on LCV-110, Olympus, but the optical system including bilateral telecentric lenses was newly designed; for obtaining the whole aggregate view, a ×0.5 lens was interposed in the light path; Supplementary Fig. 2f). For high-resolution multi-photon live imaging, a stack of optical section images (512 × 512 pixels for the X–Y plane and 2 μm for Z-axis step; typically 150 sections) was taken at each time point using a ×25 water-immersion lens (N.A. 1.05, Olympus) and multi-photon femtosecond laser (900 nm; Mai-Tai DeepSee eHP, Spectra-Physics) with group velocity dispersion auto-compensation (see Supplementary Fig. 2a, b). The 3D image analyses were done with Imaris 7.1 (Bitplane). The bright-field view in parallel to multi-photon imaging was obtained by scanning with a 900-nm femtosecond laser and detecting the transmission light with a separate photo-multiplier. Live imaging detected no substantial tangential migration or mixing of cells across different domains (non-retinal neuroectoderm, RPE and neural retina) of the ES-cell-derived epithelium, at least after day 3 (data not shown), indicating that distinct retinal domains are likely to be formed by region-specific patterning in the retinal epithelium rather than by the sorting of distinct cell types, which has been indicated as the mechanism in zebrafish retinogenesis[7].

For membrane staining, cells were stained with the vital fluorescent dye Cell Mask. For high-resolution cell-shape images and cell-tracking images (Figs 3a and 5a), cells were partially labelled by co-aggregating *Rx-GFP* and non-labelled ES cells on day 0.

Culture of isolated optic vesicles and neural retina. For excision assays, Rx–GFP^{+} optic vesicles were mechanically excised (with or without non-retinal neuroectodermal epithelium tissues) from the main body of day 7 SFEB aggregates using no. 5 forceps, and were subjected to suspension culture under 40%-O_2/5%-CO_2 conditions in DMEM/F12 medium supplemented with N2. For the explant samples with neuroectodermal epithelium, the culture included approximately the same amount of neuroectodermal epithelium tissue with Rx^{+} vesicle. For co-culture assays, Wnt3a-expressing or control L cells (both lines were purchased from ATCC) were spun down in 1.5-ml tubes and the cell pellets were entrapped in a droplet of undiluted matrigel and subjected to co-culture with the isolated vesicles without neuroectodermal epithelium tissues. In this co-culture, a small gap (100–150 μm) was placed between an L cell aggregate and an isolated vesicle to minimize physical interactions. No substantial differences in mitosis (pH3^{+}) or apoptosis (active caspase 3^{+}) were seen between the proximal epithelia of the excised and non-excised vesicles. Unlike the cup in Fig. 5c, the optic cup forming from the isolated vesicles did not evert, presumably because the proximal end of the RPE spontaneously sealed and became immobilized itself.

For long-term neural retina culture, the neural retina portions of day 10 optic cups with some RPE were isolated mechanically using forceps and were subjected to suspension culture under 40%-O_2/5%-CO_2 conditions in DMEM/F12 medium supplemented with N2 + 10% FBS + 0.5 μM all-*trans* retinoic acid (RA) + 1 mM L-taurine. The attaching RPE tissue everted or rolled up at one end of the growing neural retina vesicle during suspension culture. The culture grew and survived better with the 40% O_2 condition than with the 20% air. In particular, the survival of ganglion cells, which are first-born neurons, was much better supported by the high-O_2 condition. RA also improved the tissue growth. In the absence of RA, the photoreceptor layer tended to show some disorder in tissue architecture and contained some rosettes. In contrast, the absence of L-taurine had only a marginal effect in this culture. The neural retina culture could be continued until day 35, and, beyond this, the tissue gradually lost its integrity (in future investigation, it is necessary to establish improved conditions for longer culture to induce the further maturation of functional photoreceptors, for example, in the development of outer segments).

Electroporation of the *pNrl-DsRed2* plasmid (from Addgene) with the *pCAG-H2B-GFP* plasmid was performed on the apical side of day-16 neural retina tissues using the CUY21 pulse generator (BEX; 30 V, 50 ms × 5 pulses, 450-ms interval) and a pair of platinum electrodes (2.5-mm gap; BEX). On days 18–20, 20–30% of H2B–GFP^{+} cells co-expressed DsRed2.

Three-dimensional multi-photon laser ablation and mitosis inhibition. Three-dimensional laser ablation was done using the same multi-photon optical system for 3D live imaging. 900-nm laser beam from Mai-Tai Deep-See eHP at full power was condensed at the target cells via a ×25 water-immersion lens (N.A. 1.05) using a spot illumination or line scan until cell rupture (typically 0.4–2 s duration). The gap created by cell rupture was quantified as a maximal gapped area in the optical section at each time point. When the gap was closed, the content of ruptured cells (debris) was usually pushed out from the epithelium via the apical side, probably because the other side was blocked by the basement membrane. Although the created gaps remained open in the tissues with pulling tension (for example, the epithelia at the early phases) for a substantial period after ablation, they eventually started closing, for example, 1 h after, probably reflecting a secondary wound healing.

The mitotic inhibitor aphidicolin was used to attenuate the proliferation during phase 4 by adding it at a final concentration of 5 μM to culture at the end of phase 3 and incubating the culture for 5 h before analysis (its effects on cell proliferation and tissue expansion are shown in Supplementary Fig. 2h–j).

Immunohistochemistry, qPCR and FACS. Immunohistochemistry was performed as described previously[20]. Antibodies against the following proteins were used at the indicated dilutions: Nanog (rabbit/1:500, Reprocell), Oct3 (mouse/1:500, Transduction Laboratories), N-cadherin (mouse/1:1,000, BD Transduction), Pax6 (mouse/1:1,000, R&D), Chx10 (sheep/1:1,000, Exalpha), Rx (rabbit/1:1,000[15]), laminin (rat/1:500, Chemicon, rabbit/1:500, Abcam), PKCζ (rabbit/1:200, Santa Cruz), CD133 (rat/1:500, Chemicon), Mitf (mouse/1:1,000, Exalpha), Coup-TF2 (mouse/1:1,000, Perseus Proteomics), connexion 43 (rabbit/1:1,000, Sigma), recoverin (rabbit/1:1,000, Chemicon), Otx2 (goat/1:100, Santa Cruz), Brn3 (goat/1:50, Santa Cruz, C-13), rhodopsin (mouse/1:1,000, Sigma), calbindin (mouse/1:1,000, Swant), calretinin (rabbit/1:1,000, Swant), CRALBP (also called Rlbp1; mouse/1:500, Abcam), VGLUT1 (rabbit/1:1,000, Wako), bassoon (mouse/1:200, Assay Designs), pMLC2 (rabbit/1:500, Cell Signaling), pH3 (mouse/1:500, Cell Signaling). DAPI was used for counterstaining the nuclei (Molecular Probes). Stained sections were analysed with an LSM710 confocal microscope (Zeiss).

Quantitative PCR was performed using the 7500 Fast Real Time PCR System (Applied Biosystems) and data were normalized to GAPDH expression. Primers used were as follows: *Dkk1*, forward 5′-CCGGGAACTACTGCAAAAAT-3′, reverse 5′-CCAAGGTTTTCAATGATGCTT-3′; *Dkk3*, forward 5′-TCGTGACCAGATCCAGCTT-3′, reverse 5′-AGCCGCTGCATGTTTGTT-3′; *Rx*, forward 5′-CGACGTTCACCACTTACCAA-3′, reverse 5′-TCGGTTCTGGAACCATACCT-3′; *Six3*, forward 5′-CCGGAAGAGTTGTCCATGTTC-3′, reverse 5′-CGACTCGTGTTTGTTGATGGC-3′; *Ncad*, forward 5′-CAGGGTGGACGTCATTGTAG-3′, reverse 5′-AGGGTCTCCACCACTGATTC-3′. The values shown on graphs represent the mean ± s.e.m. For FACS analysis, cells were counted with FACSAria (Becton Dickinson), and the data were analysed with the FACSDiva software (Becton Dickinson).

Birth-dating analysis. For the *in vitro* birth-dating analysis, aggregates were treated with BrdU (5 μg ml^{-1}, Invitrogen) on days 10, 12, 14, 16, 18 or 20, and

©2011 Macmillan Publishers Limited. All rights reserved

rinsed with medium 12 h after, to remove the BrdU. On day 24, the cell aggregates were fixed and cryosectioned. Sections were immunostained and the percentages of BrdU^{+} cells that were positive for each lineage marker were quantified. For quantification, 16–24 aggregates were examined for each experiment, which was repeated at least three times.

Tissue elasticity measurement. The tissue elasticity/rigidity of ES-cell-derived retinal epithelia was examined by indentation assay using an AFM cantilever as described previously[35,36]. Force probe was prepared by attaching a glass bead (100 μm diameter) to a tipless silicon cantilever (450 μm long, 50 μm wide, 2 μm thick; nominal spring constant ~0.02 N m^{-1}; TL-CONT, Nanosensors) using two-component Araldite epoxy glue (the spring constant of the cantilever with bead was typically ~0.1 N m^{-1}). The measurement using a contact mode was carried out with the Nano-Wizard system with the Cell-Hesion module (JPK), which is optimized for biological samples. Before the measurement, the spring constant of the force probe was calibrated according to the manufacturer's instruction by the thermal fluctuation method. The retinal tissues were pressed with a force of 20 nN (the approach and retraction speeds were set to 10–20 μm s^{-1}) by the force probe and the data analysis was done with the JPK image processing software, in which the tissue rigidity is calculated as a value of pushing stress (Pa) necessary for unitary strain of the tissue (Young's modulus determined by a force–distance curve). The measurement was done on the apical side of ES-cell-derived retinal tissues in the culture medium (DMEM/F12/N2) on top of a poly-D-lysine-coated plastic culture dish (several points in each RPE or neural retina epithelium). The inhibitor treatments were given for 60 min at 37 °C before the measurements. All measurements were performed at room temperature.

Computer-simulated animations for the working model. A two-dimensional continuum model of epithelium consisting of multiple cells along the tangential direction was discretized into quadrilateral elements expressed by their vertex coordinates $\mathbf{r}_i$ as shown in Supplementary Fig. 6b. The dynamic formation process of the optic cup structure was numerically demonstrated under the axisymmetric condition to reduce the geometric complexity.

The vertices were connected by linear springs representing elastic behaviours of the element sides, and their motions were determined to decrease the total potential energy of the system U:

$$U = U_e + U_v + U_s \tag{1}$$

by solving the overdamped equation of motion[51–53]:

$$\eta \frac{d\mathbf{r}_i}{dt} = -\frac{\partial U}{\partial \mathbf{r}_i} \tag{2}$$

where η represents the viscosity of the system. The energy U_e is elastic strain energy of the springs expressed as

$$U_e = \sum \frac{1}{2} k_e^{\alpha} \left(l^{\alpha} - l_{eq}^{\alpha} \right)^2 \tag{3}$$

where l^{α} is the length of element side, l_{eq}^{α} is its natural length at the stress-free state, and k_e^{α} is the elastic constant for $\alpha = a$ (apical), b (basal), t (transmural). The energy U_v for volumetric change of the elements

$$U_v = \sum \frac{1}{2} k_v \left(V - V_{eq} \right)^2 \tag{4}$$

was introduced to take into account the local volumetric constraint of the elements, where V is the element volume, V_{eq} is the equilibrium element volume, and k_v is the volumetric elastic constant. The energy U_s was introduced to express the resistance to the distortion of the elements (the 'self-righting' effect), as

$$U_s = \sum \frac{1}{2} k_s \left\{ 1 - \cos\left(\theta_1^{\alpha} - \theta_2^{\alpha} \right) \right\} \tag{5}$$

where θ_1^{α} and θ_2^{α} are the angles of element corners for $\alpha = a$ (apical), b (basal), and k_s is the elastic constant.

A hemispherical model of the epithelium vesicle with a constant thickness in the transmural direction was initially created in the stress-free state as a reference configuration for the elastic energies, and the epithelium, from the neuroectodermal epithelium part to the centre of the neural retina, was equally discretized into 100 elements. Deformation of the epithelium in the r–z plane generated by the hoop stress (circumferential stress), which comes from the three-dimensionality (curvature in the circumferential direction) of the vesicle under the axisymmetric condition, was considered by applying apparent force to each vertex in the radial direction that was assumed to be simply proportional to the magnitude of radial displacement of the vertex referring to its stress-free configuration. Then, a half domain of the hemispherical epithelium vesicle was analysed by assuming the symmetry with respect to the z axis as shown in Supplementary Fig. 6b.

The computer-simulated animations were produced on the basis of the following assumption deduced from the experimental results of the present study: (1) relaxation of the presumptive neural retina beginning at phase 2; (2) strong apical constriction of the hinge domain at phase 3; and (3) rapid tangential growth of the distal RPE and (tangential/vertical) growth of the neural retina at phase 4.

Phase 1: to create the initial condition, apical contraction in the retinal epithelium was simulated by reducing the natural length of the elements l_{eq}^{a}, which results in the formation of the evaginating vesicular shape.

Phase 2: elastic strain energy U_e stored in the tangential elements on the apical and basal surfaces in the presumptive neural retina was gradually released by changing their natural length, l_{eq}^{a} and l_{eq}^{b}, to the relaxed configurations with decreased elastic constants k_v and k_s. Simultaneously, equilibrium element volume V_{eq} in the neural retina region was relaxed to the corresponding stress-free state.

Phase 3: strong apical constriction in the hinge epithelium was simulated by reducing the natural length l_{eq}^{a} with an increase in the elastic constant ${k_e}^{a}$ at the corresponding region. In the neural retina, the apical and basal elastic constants (${k_e}^{a}$ and ${k_e}^{b}$) were gradually decreased until the end of phase 3 (by 75% and also kept at this level during phase 4).

Phase 4: tangential expansion of the distal RPE and the neural retina at constant growth rates was simulated by increasing the size (natural length and equilibrium volume) of elements in the corresponding directions. To mimic the *in vitro* morphology, vertical expansion was also introduced in the central portion of the neural retina during phases 3–4 (although this did not show a strong impact on the neural retina invagination *in silico*).

51. Honda, H., Tanemura, M. & Nagai, T. A three-dimensional vertex dynamics cell model of space-filling polyhedra simulating cell behaviour in a cell aggregate. *J. Theor. Biol.* **226,** 439–453 (2004).
52. Nagai, T. & Honda, H. Computer simulation of wound closure in epithelial tissues: cell-basal-lamina adhesion. *Phys. Rev. E* **80,** 061903 (2009).
53. Inoue, Y. & Adachi, T. Coarse-grained Brownian ratchet model of membrane protrusion on cellular scale. *Biomech. Model. Mechanobiol.* doi:10.1007/s10237-010-0250-6 (19 August 2010).

©2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09922

Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease

Zeneng Wang[1,2], Elizabeth Klipfell[1,2], Brian J. Bennett[3], Robert Koeth[1], Bruce S. Levison[1,2], Brandon DuGar[1], Ariel E. Feldstein[1,2], Earl B. Britt[1,2], Xiaoming Fu[1,2], Yoon-Mi Chung[1,2], Yuping Wu[4], Phil Schauer[5], Jonathan D. Smith[1,6], Hooman Allayee[7], W. H. Wilson Tang[1,2,6], Joseph A. DiDonato[1,2], Aldons J. Lusis[3] & Stanley L. Hazen[1,2,6]

Metabolomics studies hold promise for the discovery of pathways linked to disease processes. Cardiovascular disease (CVD) represents the leading cause of death and morbidity worldwide. Here we used a metabolomics approach to generate unbiased small-molecule metabolic profiles in plasma that predict risk for CVD. Three metabolites of the dietary lipid phosphatidylcholine—choline, trimethylamine *N*-oxide (TMAO) and betaine—were identified and then shown to predict risk for CVD in an independent large clinical cohort. Dietary supplementation of mice with choline, TMAO or betaine promoted upregulation of multiple macrophage scavenger receptors linked to atherosclerosis, and supplementation with choline or TMAO promoted atherosclerosis. Studies using germ-free mice confirmed a critical role for dietary choline and gut flora in TMAO production, augmented macrophage cholesterol accumulation and foam cell formation. Suppression of intestinal microflora in atherosclerosis-prone mice inhibited dietary-choline-enhanced atherosclerosis. Genetic variations controlling expression of flavin monooxygenases, an enzymatic source of TMAO, segregated with atherosclerosis in hyperlipidaemic mice. Discovery of a relationship between gut-flora-dependent metabolism of dietary phosphatidylcholine and CVD pathogenesis provides opportunities for the development of new diagnostic tests and therapeutic approaches for atherosclerotic heart disease.

The pathogenesis of CVD includes genetic and environmental factors. A known environmental risk factor for the development of CVD is a diet rich in lipids. A relationship between blood cholesterol and triglyceride levels and cardiovascular risk is well established. However, less is known about the role of the third major category of lipids, phospholipids, in atherosclerotic heart disease pathogenesis.

Another potential yet controversial environmental factor in the development or progression of atherosclerotic heart disease is inflammation due to infectious agents. Some studies have reported associations between coronary disease and pathogens such as cytomegalovirus (CMV), *Helicobactor pylori*, and *Chlamydia pneumoniae*[1–4]. However, prospective randomized trials with antibiotics in humans have thus far failed to demonstrate cardiovascular benefit[5–7] and studies with germ-free hyperlipidaemic mice confirm that infectious agents are not necessary for murine atherosclerotic plaque development[8]. Although a definite cause-and-effect relationship between a bacterial or viral pathogen and atherosclerosis in humans has not yet been established, the prospect of a role for microbes in atherosclerosis susceptibility remains enticing.

The intestinal microbiota ('gut flora'), comprised of trillions of typically non-pathogenic commensal organisms, serve as a filter for our greatest environmental exposure—what we eat. Gut flora have an essential role, aiding in the digestion and absorption of many nutrients[9]. Animal studies have recently shown that intestinal microbial communities can influence the efficiency of harvesting energy from diet, and consequently influence susceptibility to obesity[10]. Metabolomics studies of inbred mouse strains have also recently shown that gut microbiota may have an active role in the development of complex metabolic abnormalities, such as susceptibility to insulin resistance and non-alcoholic fatty liver disease[11]. A link between gut-flora-dependent phospholipid metabolism and atherosclerosis risk through generation of pro-atherosclerotic metabolites has not yet been reported.

Metabolomics studies identify markers of CVD

In initial studies we sought to discover unbiased small-molecule metabolic profiles in plasma that predict increased risk for CVD. An initial 'Learning Cohort' was used comprising plasma from stable patients undergoing elective cardiac evaluation who subsequently experienced a heart attack (myocardial infarction), stroke or death over the ensuing three-year period versus age- and gender-matched subjects who did not. Liquid chromatography with on-line mass spectrometry (LC/MS) analysis of plasma was performed to define analytes associated with cardiac risk as described in Methods. Of an initial 2,000+ analytes monitored, 40 met all acceptability criteria within the Learning Cohort. Subsequent studies within an independent 'Validation Cohort' led to identification of 18 analytes that met acceptability criteria in both Learning and Validation Cohorts (Fig. 1a, b, Supplementary Fig. 1a and Supplementary Table 1).

The structural identity of the 18 small molecules in plasma, the levels of which track with cardiac risks, was not known, as the compounds were screened on the basis of retention time and mass-to-charge ratio (m/z) when analysed by LC/MS. Among the 18 analytes, those with m/z 76, 104 and 118 demonstrated significant ($P < 0.001$) correlations among one another, suggesting a potential relationship via a common biochemical pathway (Supplementary Fig. 1b). We therefore initially sought to structurally define these three analytes.

Phosphatidylcholine metabolites are linked to CVD

The candidate compound in plasma with an m/z of 76 associated with CVD risks was isolated and unambiguously identified as TMAO using

[1]Department of Cell Biology, Cleveland Clinic, Cleveland, Ohio 44195, USA. [2]Center for Cardiovascular Diagnostics and Prevention, Cleveland Clinic, Cleveland, Ohio 44195, USA. [3]Department of Medicine/Division of Cardiology, BH-307 Center for the Health Sciences, University of California, Los Angeles, California 90095, USA. [4]Department of Mathematics, Cleveland State University, Cleveland, Ohio 44115, USA. [5]Bariatric and Metabolic Institute, Cleveland Clinic, Cleveland, Ohio 44195, USA. [6]Department of Cardiovascular Medicine, Cleveland Clinic, Cleveland, Ohio 44195, USA. [7]Department of Preventive Medicine and Institute for Genetic Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90089, USA.

©2011 Macmillan Publishers Limited. All rights reserved

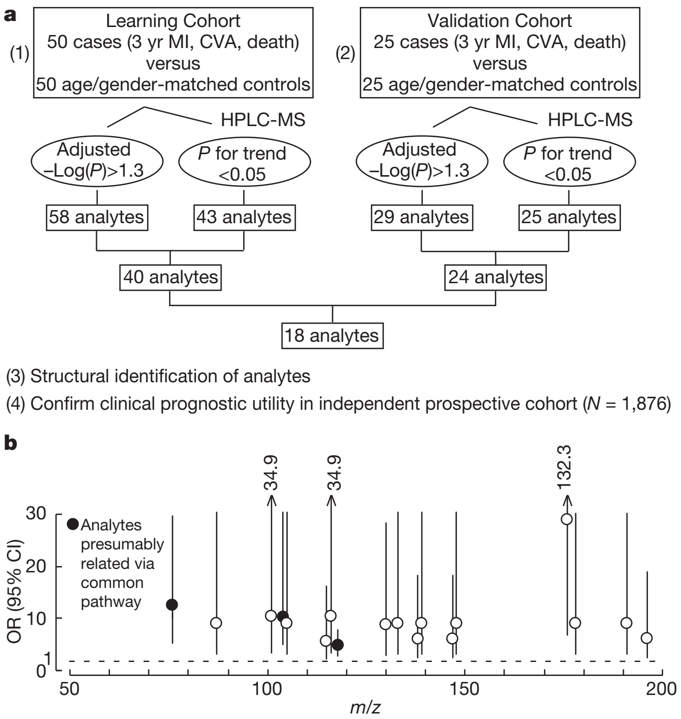


Figure 1 | Strategy for metabolomics studies to identify plasma analytes associated with cardiovascular risk. a, Overall schematic to identify plasma analytes associated with cardiac risk over the ensuing 3-year period. CVA, cerebrovascular accident; HPLC, high-performance liquid chromatography; MI, myocardial infarction. **b**, Odds ratio (OR) and 95% confidence intervals (CI) of incident (3-year) risk for MI, CVA or death of the 18 plasma analytes that met all selection criteria in both Learning and Validation Cohorts; odds ratio and 95% confidence intervals shown are for the highest versus lowest quartile for each analyte. Filled circles represent the analytes (*m/z* = 76, 104, 118) focused on in this study. *m/z*, mass to charge ratio.

multinuclear nuclear magnetic resonance (NMR), multi-stage mass spectrometry (MS^n), liquid chromatography with tandem mass spectrometry (LC/MS/MS) and gas chromatography with tandem mass spectrometry (GC/MS/MS) after multiple derivatization strategies (see Methods, Supplementary Figs 2a–d, and Supplementary Table 2). TMAO, an oxidation product of trimethylamine (TMA), is a relatively common metabolite of choline in animals[12,13]. Foods rich in the lipid phosphatidylcholine (PC, also called lecithin), which predominantly include eggs, milk, liver, red meat, poultry, shell fish and fish, are believed to be the major dietary sources for choline, and hence TMAO production[14]. Briefly, initial catabolism of choline and other trimethylamine-containing species (for example, betaine) by intestinal microbes forms the gas TMA[13], which is efficiently absorbed and rapidly metabolized by at least one member of the hepatic flavin monooxygenase (FMO) family of enzymes, FMO3, to form TMAO[15,16]. Identification of the plasma analyte associated with CVD risk with an *m/z* of 76 as TMAO therefore indicated that the plasma analyte with an *m/z* of 104 might be choline. Further, these results also indicated that the plasma analyte with an *m/z* of 118 associated with CVD might be related to PC (choline) metabolism.

To test the hypothesis that the plasma analytes with *m/z* 76 (TMAO), 104 and 118 might all be derived from the major dietary lipid PC, mice were fed egg-yolk PC (through oral gavage) and plasma levels of analytes over time were monitored. In both male and female mice, analytes with the same *m/z* (76, 104 and 118) and the same retention times as the corresponding analytes of interest observed in human plasma all increased after oral PC feeding (Supplementary Fig. 3a, b), strongly indicating that the *m/z* 104 analyte was choline, and the analyte at *m/z* 118 was derived from PC. Confirmation that the plasma analyte (*m/z* 104) associated with CVD risk was choline was achieved by MS^n, LC/MS/MS and GC/MS/MS after multiple derivitization strategies (Supplementary Fig. 4a–d and Supplementary Table 3).

We next studied the plasma analyte with *m/z* 118. We proposed that the analyte was either betaine or one of several potential methylated metabolites of choline (see Supplementary Fig. 5a for structures and strategy for discrimination among these isomers). To distinguish between these species, and explore a role for intestinal generation of the various metabolites, different isotopically labelled choline precursors were administered to mice either through an oral (gavage) or a parenteral (intraperitoneal, i.p.) route. The observed *m/z* of new isotopically labelled analytes at the appropriate retention times identified in plasma after these isotope tracer studies are summarized in Fig. 2a. Oral administration of non-labelled choline resulted in time-dependent increases in plasma levels of analytes with *m/z* 76, 104 and 118, consistent with TMAO, choline and either betaine or a methylated choline species (Supplementary Fig. 6a). Use of selectively deuterated choline species at either the trimethylamine moiety (d9 isotopomer) or the ethyl moiety (d4 isotopomer) unambiguously confirmed the *m/z* 118 analyte as betaine (Fig. 2a and Supplementary Fig. 6b). Further confirmation was acquired by observing the same retention time in LC/MS and an identical collision-induced dissociation (CID) mass spectrum (Supplementary Fig. 5b). Moreover, supplementation of PC or choline isotopomers via gavage or i.p. injection showed an absolute requirement for the oral route in TMAO production, whereas betaine production from PC or choline was formed via both oral and i.p. routes (Fig. 2 and Supplementary Fig. 7a).

Gut flora is needed to form TMAO from dietary PC

Intestinal microflora have a role in TMAO formation from dietary free choline[13]. We therefore proposed that commensal organisms (gut flora) might also have an obligate role in TMAO formation from dietary PC. To test this, deuterated PC was synthesized whereby the choline-methyl groups were deuterium labelled (that is, d9-PC) and

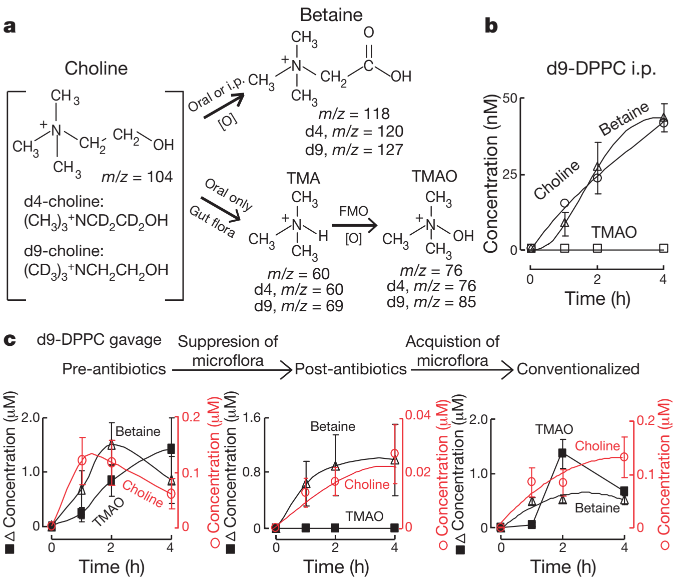


Figure 2 | Identification of metabolites of dietary PC and an obligatory role for gut flora in generation of plasma analytes associated with CVD risks. a, Summary schematic indicating structure of metabolites and routes (oral or i.p.) of formation observed in choline challenge studies in mice using the indicated isotope-labelled choline. The *m/z* in plasma observed for the isotopomers of the choline metabolites are shown. **b**, Plasma levels of d9 metabolites after i.p. challenge with d9(trimethyl)-dipalmitoylphosphatidylcholine (d9-DPPC). **c**, d9-TMAO production after oral d9-DPPC administration in mice, following suppression of gut flora with antibiotics (3 weeks), and then following placement (4 weeks) into conventional cages with non-sterile mice ('conventionalized'). Data are presented as mean ± standard error (s.e.) from four independent replicates.

©2011 Macmillan Publishers Limited. All rights reserved

used as isotope tracer for feeding studies. When mice were fed through oral gavage with d9-PC, the time-dependent appearance of the anticipated d9 isotopomer of TMAO was observed in plasma (Fig. 2c). Interestingly, pre-treatment of mice with a three-week course of broad-spectrum antibiotics to suppress intestinal flora completely suppressed the appearance of d9-TMAO in plasma after oral d9-PC administration (Fig. 2c). A similar pattern was observed after oral administration of d9-choline to mice, with d9-TMAO produced in untreated mice, but not in the same mice after a 3-week course of broad-spectrum antibiotics (Supplementary Fig. 7b), or in germ-free mice born sterilely by Caesarean section (Supplementary Fig. 7c). In a final series of studies, mice with suppressed intestinal microflora after antibiotics were placed in conventional cages with normal (non-germ-free) mice to permit intestinal colonization with microbes. After four weeks, repeat oral d9-PC challenge of the now 'conventionalized' mice resulted in readily detectable plasma levels of d9-TMAO (Fig. 2c). Similar results were observed after conventionalization of germ-free mice and oral d9-choline (Supplementary Fig. 7c). Collectively, these results show an obligate role for intestinal microbiota in the generation of TMAO from the dietary lipid PC. They also reveal the following metabolic pathway for dietary PC producing TMAO: PC → choline → TMA → TMAO.

Dietary PC metabolites predict CVD risk

We next sought to independently confirm the prognostic value of monitoring fasting plasma levels of TMAO, choline and betaine in a large independent clinical cohort distinct from subjects examined in both the Learning and Validation Cohorts. Stable subjects ($N = 1{,}876$) undergoing elective cardiac evaluations were enrolled. Clinical, demographic and laboratory characteristics of the cohort are provided in Supplementary Table 4a. Fasting plasma levels of TMAO, choline and betaine were quantified by LC/MS/MS using methods specific for each analyte (Supplementary Fig. 8). Increasing levels of choline, TMAO and betaine were all observed to show dose-dependent associations with the presence of CVD (Fig. 3a–c) and multiple individual CVD phenotypes including peripheral artery disease (PAD), coronary artery disease (CAD), and history of myocardial infarction (see Supplementary Table 5a–d for multilogistic regression models, and Supplementary Table 5e for Somers' Dxy correlation). The associations between increased risk of all CVD phenotypes monitored and elevated systemic levels of the three PC metabolites held true after adjustments for traditional cardiac risk factors and medication usage (Fig. 3a–c and Supplementary Table 5a–e).

Dietary choline or TMAO promotes atherosclerosis

We next investigated whether the strong associations noted between plasma levels of the dietary PC metabolites and CVD risk reflected some hidden underlying pro-atherosclerotic mechanism. Atherosclerosis-prone mice (C57BL/6J *Apoe*$^{-/-}$) at time of weaning were placed on either normal chow diet (contains 0.08–0.09% total choline, wt/wt) or normal chow diet supplemented with intermediate (0.5%) or high amounts of additional choline (1.0%) or TMAO (0.12%). At 20 weeks of age increased total aortic root atherosclerotic plaque area was noted in both male and female mice on diets supplemented with either choline or TMAO (Fig. 3d and Supplementary Fig. 9a). Analysis of plasma levels of choline and TMAO in each of the dietary arms showed nominal changes in plasma levels of choline, but significant increases of TMAO in mice receiving either choline or TMAO supplementation (Supplementary Fig. 10). Parallel examination of plasma cholesterol, triglycerides, lipoproteins, glucose levels and hepatic triglyceride content in the mice failed to show significant increases that could account for the enhanced atherosclerosis (Supplementary Table 6 and Supplementary Fig. 11). Interestingly, all dietary groups of mice revealed a significant positive correlation between plasma levels of TMAO and atherosclerotic plaque size (Fig. 3e and Supplementary Fig. 9b). Of note, plasma TMAO levels observed within the female mice (which get enhanced atherosclerosis relative to their male counterparts), even on normal chow diet, were significantly higher than those observed among male mice (Supplementary Fig. 10). No significant gender differences in plasma levels of TMAO were observed in humans ($P = 0.47$); however, a clear dose–response relationship was observed between TMAO levels and clinical atherosclerotic plaque burden in subjects undergoing coronary angiography (Fig. 3f).

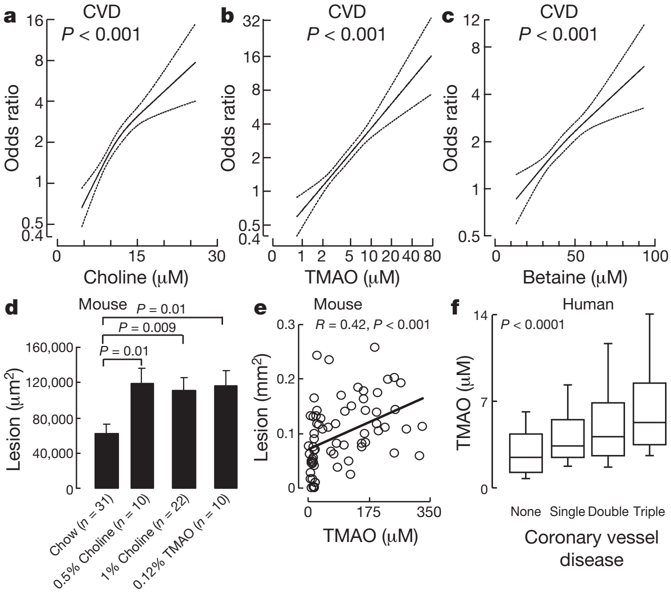


Figure 3 | Plasma levels of choline, TMAO and betaine are associated with atherosclerosis risks in humans and promote atherosclerosis in mice. **a**–**c**, Spline models of the logistic regression analyses reflecting risk of CVD (with 95% CI) according to plasma levels of choline, TMAO and betaine in the entire cohort ($n = 1{,}876$ subjects). **d**, Comparison in aortic lesion area among 20-week-old female C57BL/6J *Apoe*$^{-/-}$ mice fed with chow diet supplemented with the indicated amounts (wt/wt) of choline or TMAO from time of weaning (4 weeks). **e**, Relationship between plasma TMAO levels and aortic lesion area. **f**, Relationship between fasting plasma levels of TMAO versus CAD burden among subjects ($N = 1{,}020$). Boxes represent 25th, 50th and 75th percentile, and whiskers 5th and 95th percentile plasma levels. Single, double and triple coronary vessel disease refers to number of major coronary vessels demonstrating ≥50% stenosis on diagnostic coronary angiography.

Hepatic FMOs, TMAO and atherosclerosis

Hepatic FMO3 is a known enzymatic source for TMAO in humans, based on the recent recognition of the aetiology of an uncommon genetic disorder called trimethylaminuria (also known as fish malodour syndrome)[15,17]. Subjects with this metabolic condition have impaired capacity to convert TMA, which smells like rotting fish, into TMAO, an odourless stable oxidation product[17]. We therefore sought to identify possible sources of genetic regulation and the role of *Fmo3* in atherosclerosis using integrative genetics in mice[18]. Expression levels of *Fmo3* were determined by microarray analysis in the livers of mice from an F2 intercross between atherosclerosis-prone C57BL/6J *Apoe*$^{-/-}$ mice and atherosclerosis-resistant C3H/HeJ *Apoe*$^{-/-}$ mice and compared with quantitative measures of atherosclerosis. The expression level of *Fmo3* showed marked differences between genders (females >1,000 fold higher than in males). Significant positive correlations were readily found between hepatic *Fmo3* expression and atherosclerotic lesions (Fig. 4a, Supplementary Fig. 12, top row, and Supplementary Fig. 13). Interestingly, a highly significant negative correlation with plasma high-density lipoprotein (HDL) cholesterol levels was noted in both male and female mice (Fig. 4b and Supplementary Fig. 12, middle row). Further, plasma levels of the PC metabolite TMAO showed a significant positive correlation with hepatic *Fmo3* expression level in mice (Fig. 4c and Supplementary Fig. 12, bottom row).

©2011 Macmillan Publishers Limited. All rights reserved

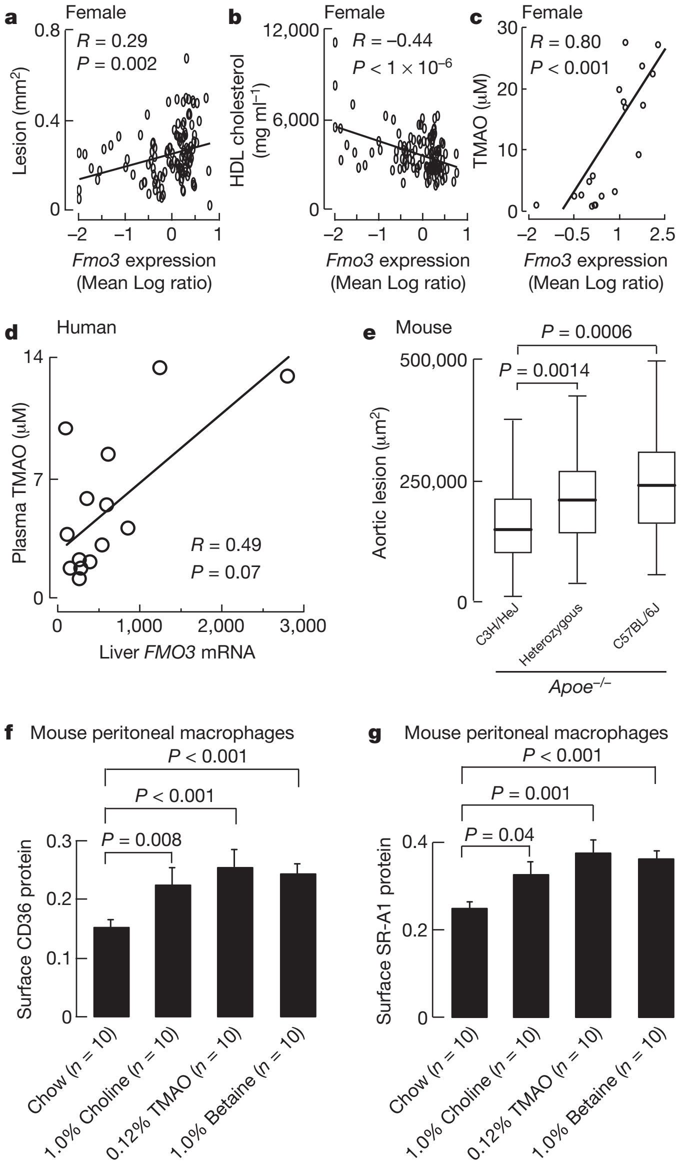


Figure 4 | Hepatic *Fmo* genes are linked to atherosclerosis and dietary PC metabolites enhance macrophage scavenger receptor expression. **a**–**c**, Correlation between hepatic *Fmo3* expression and aortic lesion, plasma HDL cholesterol and TMAO in female mice from the F2 intercross between atherosclerosis-prone C57BL/6J *Apoe*$^{-/-}$ and atherosclerosis-resistant C3H/HeJ *Apoe*$^{-/-}$ mice. **d**, Correlation between human hepatic *FMO3* expression and plasma TMAO. **e**, Effect of *Fmo3* genotype (SNP rs3689151) on aortic sinus atherosclerosis in male mice from the C57BL6/J *Apoe*$^{-/-}$ and C3H/HeJ *Apoe*$^{-/-}$ F2 intercross. **f**, **g**, Quantification of scavenger receptor CD36 and SR-A1 surface protein levels in macrophages harvested from C57BL/6J mice (13 week) after three weeks of standard chow versus chow supplemented with the indicated amounts (wt/wt) of choline, TMAO or betaine. Data are presented as mean ± s.e. from the indicated numbers of mice in each group.

FMO3 is one member of a family of FMO enzymes, the majority of which are physically located as a cluster of genes on chromosome 1 in both humans and mice. The various FMOs share sequence homology and overlapping substrate specificities. Further, although rare mutations in or near the *FMO3* gene have been identified in individuals with trimethylaminuria[19], the impact of these mutations on other *FMO* genes remains unknown. Examination of the hepatic expression levels of the various *FMO* genes revealed that many are highly correlated with each other in both mice and humans (Supplementary Table 7). Examination of hepatic expression levels of additional *Fmo* genes in mice from the atherosclerosis F2 intercross revealed that multiple *Fmo* genes are significantly correlated with aortic lesion formation, HDL cholesterol concentrations and plasma TMAO levels (Supplementary Figs 14–16), suggesting that several members of the FMO family of enzymes may participate in atherosclerosis and the PC → TMAO metabolic pathway. To explore the relationship between hepatic FMOs and plasma TMAO levels in humans, paired samples of liver and plasma from subjects undergoing elective liver biopsy were examined. Among all of the human *FMO* genes monitored, only a trend towards positive association was noted between hepatic expression of *FMO3* and plasma TMAO levels (Fig. 4d and Supplementary Fig. 17).

Next, we focused on the genetic regulation of hepatic *Fmo3* expression (and other *Fmo* genes) using expression quantitative trait locus (eQTL) analyses in the F2 mouse intercross. The eQTL plot for *Fmo3* messenger RNA levels is shown in Supplementary Fig. 18, and demonstrates a strongly suggestive *cis* locus (lod score = 5.9) on mouse chromosome 1 at 151 Mb. *Fmo3* (and several other *Fmo* genes) is located at 164.8 Mb in a region identified as non-identical by descent between C3H/HeJ and C57BL/6 (http://mouse.cs.ucla.edu/perlegen/). This region is just distal to the 95% confidence interval of a previously reported murine atherosclerosis susceptibility locus[20]. Examining the effect of the closest single-nucleotide polymorphism (SNP) to *Fmo3* (rs3689151) as a function of alleles inherited from either parental strain indicated a strong effect on atherosclerosis in both genders of the F2 mice (Kruskal–Wallis test, $P < 1.0 \times 10^{-6}$). Bonferroni corrected pairwise comparisons indicated a dose-dependent significant increase in atherosclerosis in F2 mice heterozygous or homozygous for the C57BL/6J allele (Fig. 4e). Although the resolution on average for an F2 intercross of this size is in excess of 20 Mb and thus does not provide 'gene-level' resolution, these data show that the locus encompassing the *Fmo* gene cluster on chromosome 1 is associated with atherosclerotic lesion size. Collectively, these results indicate that: (1) hepatic expression levels of multiple *Fmo* genes are linked to plasma TMAO levels in mice; (2) hepatic expression levels of multiple *Fmo* genes are associated with both the extent of aortic atherosclerosis and HDL cholesterol levels in mice; (3) hepatic expression levels of *FMO3* indicate an association with plasma TMAO levels in humans; and (4) a genetic locus containing the *Fmo* gene cluster on chromosome 1 in mice has a strong effect on atherosclerosis.

Diet and gut flora alter macrophage phenotype

To explore potential mechanisms through which dietary choline and its metabolites might exert their pro-atherosclerotic effects, C57BL/6J *Apoe*$^{-/-}$ mice at time of weaning were placed on a normal chow diet supplemented with either choline, TMAO or betaine (for >3 weeks). Both mRNA levels (Supplementary Fig. 19) and surface protein levels (Fig. 4f, g and Supplementary Fig. 20) of two macrophage scavenger receptors implicated in atherosclerosis, CD36 and SR-A1, were then determined in peritoneal macrophages. Relative to normal chow diet, mice supplemented with either choline, TMAO or betaine all showed enhanced macrophage levels of CD36 and SR-A1. We next examined the impact of dietary choline and gut flora on endogenous formation of cholesterol-laden macrophage foam cells, one of the earliest cellular hallmarks of the atherosclerotic process. Hyperlipidaemic C57BL/6J *Apoe*$^{-/-}$ mice were fed diets with defined levels of choline as follows: (1) 'control' (0.07–0.08%, wt/wt), which is similar to the choline content of normal chow (0.08–0.09%); versus (2) high 'choline', corresponding to a >10-fold higher level of choline (1.0%, wt/wt) than normal chow. Concomitantly, half of the mice were administered broad-spectrum antibiotics for 3 weeks to suppress intestinal microflora, which was confirmed by the reduction of plasma TMAO levels by >100-fold (plasma TMAO concentrations in groups receiving antibiotics were <100 nM). Whereas mice on the control diet showed modest evidence of endogenous macrophage foam cell formation, as indicated by Oil-red-O staining of peritoneal macrophages, mice on the 1% choline supplemented diet showed markedly enhanced lipid-laden macrophage development (Fig. 5a). In contrast, suppression of intestinal flora

©2011 Macmillan Publishers Limited. All rights reserved

significantly inhibited dietary-choline-induced macrophage foam cell formation (Fig. 5a, b). These results were confirmed by microscopic quantification of endogenous foam cell levels (Fig. 5b) and analytical quantification of the cholesterol content of recovered macrophages (Fig. 5c). Histopathology and biochemical studies of livers recovered from these mice showed no evidence of steatosis (Supplementary Fig. 21). Parallel analyses of plasma PC metabolites also demonstrated no significant changes in choline or betaine levels between the different dietary groups, and significant increases of plasma TMAO levels only in mice on the high-choline diet in the absence of antibiotics (males, control versus choline diet, 2.5 ± 0.1 μM versus 28.3 ± 2.4 μM, $P < 0.001$; for females, control versus choline diet, 4.0 ± 0.5 μM versus 158.6 ± 32.9 μM, $P < 0.001$).

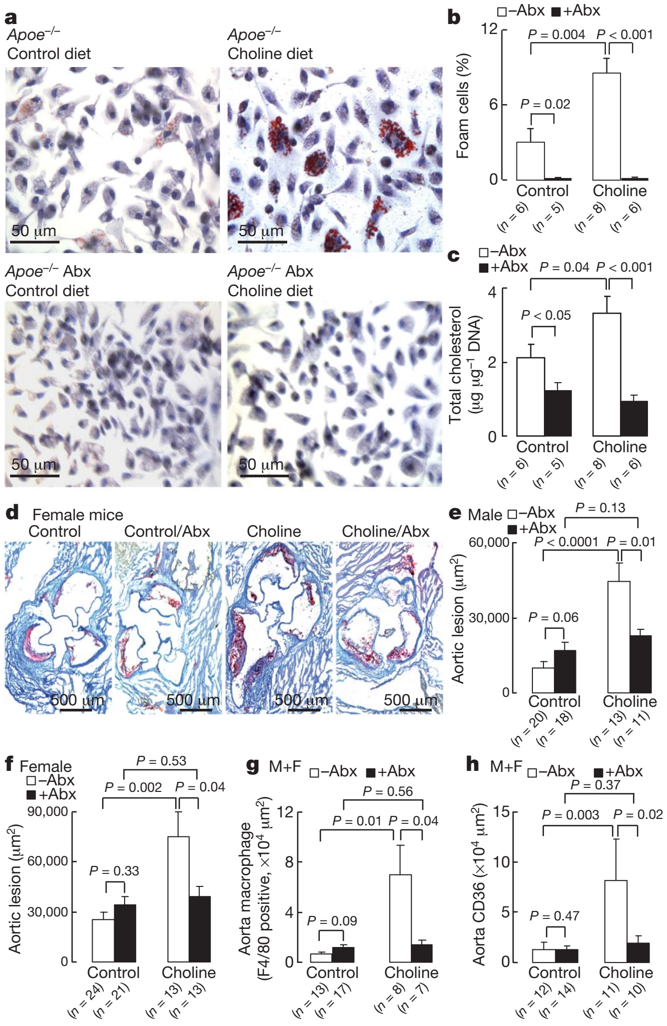


Figure 5 | Obligatory role of gut flora in dietary choline enhanced atherosclerosis. a, Choline supplementation promotes macrophage foam cell formation in a gut-flora-dependent fashion. C57BL/6J *Apoe*$^{-/-}$ mice at time of weaning (4 weeks) were provided drinking water with or without broad-spectrum antibiotics (Abx), and placed on chemically defined diets similar in composition to normal chow (control diet, 0.08 ± 0.01% total choline, wt/wt) or normal chow with high choline (choline diet, 1.00% ± 0.01% total choline, wt/wt). Resident peritoneal macrophages were recovered at 20 weeks of age. Typical images of Oil-red-O/haematoxylin-stained macrophages in each diet group are shown. **b**, Foam cell quantification from peritoneal macrophages recovered from mice in studies described in panel **a**. **c**, Macrophage cellular cholesterol content. **d**, Representative Oil-red-O/haematoxylin-stained aortic root sections from female C57BL/6J *Apoe*$^{-/-}$ mice fed control and high-choline diets in the presence or absence of antibiotics. **e**, **f**, Aortic lesion area in 20 week old C57BL/6J *Apoe*$^{-/-}$ mice off or on antibiotics and fed with control or choline diet. **g**, Aortic macrophage quantification with anti-F4/80 antibody staining. **h**, Quantification of the scavenger receptor CD36 in aorta within the indicated groups. Error bars represent s.e.m. from the indicated numbers of mice.

Gut flora promote diet-induced atherosclerosis

In additional studies we sought to test whether gut flora is involved in dietary choline-induced atherosclerosis. At the time of weaning (4 weeks old), atherosclerosis-prone C57BL/6J *Apoe*$^{-/-}$ mice were placed on either a control diet (0.08 ± 0.01%, wt/wt, choline) or a diet supplemented with 1% choline (wt/wt, choline diet). Half of the mice were also treated with broad-spectrum antibiotics to suppress intestinal microflora. Serial plasma measurements confirmed suppression of TMAO levels to virtually non-detectable levels (<100 nM) throughout the duration of the study. At 20 weeks of age, mice were killed and aortic root lesion development was quantified. In the absence of antibiotics (that is, with preserved intestinal microflora), choline supplementation augmented atherosclerosis in both male and female mice nearly three-fold (Figs 5d–f). In contrast, suppression of intestinal flora completely inhibited dietary choline-mediated enhancement in atherosclerosis (Figs 5d–f). Aortic macrophage content and scavenger receptor CD36 immunoreactive surface area within aortic lesions were markedly increased in mice on the high-choline diet, but not when intestinal microflora was suppressed with antibiotic treatment (Fig. 5g, h and Supplementary Figs 22, 23). Both histopathological and biochemical examination of liver sections from mice showed no evidence of steatosis or altered neutral lipid (triglyceride or cholesterol/cholesterol ester) levels on either diet in the absence or presence of antibiotics (Supplementary Fig. 21 and Supplementary Table 8). Finally, the structural specificity of PC metabolites in promoting a pro-atherogenic macrophage phenotype was examined. Mice fed diets supplemented with trimethylamine species (choline or TMAO) showed increased peritoneal macrophage cholesterol content and raised plasma levels of TMAO. In contrast, dietary supplementation with the choline analogue 3,3-dimethyl-1-butanol (DMB), where the quaternary amine nitrogen of choline is replaced with a carbon, resulted in no TMAO increase and no increased cholesterol in macrophages (Supplementary Fig. 24).

Discussion

Using a targeted metabolomics approach aimed at identifying plasma metabolites the levels of which predict risk of CVD in subjects, we have identified a novel pathway linking dietary lipid intake, intestinal microflora and atherosclerosis (Fig. 6). The pathway identified (dietary PC/choline → gut-flora-formed TMA → hepatic-FMO-formed TMAO) represents a unique additional nutritional contribution to the pathogenesis of CVD that involves PC and choline metabolism, an obligate role for the intestinal microbial community, and regulation of surface expression levels of macrophage scavenger receptors known to participate in the atherosclerotic process. The pro-atherogenic gut-flora-generated metabolite TMAO is formed in a two-step process initiated by gut-flora-dependent cleavage of a trimethylamine species (for example, PC, choline, betaine) generating the precursor TMA, and subsequent oxidation by FMO3 and possibly other FMOs (Fig. 6). PC is by far the most abundant dietary source of choline in most humans. The present results indicate that both environmental exposure (dietary lipid from predominantly animal products) and microbial flora participate in TMAO formation and producing a pro-atherogenic macrophage phenotype. Although the present genetic studies also indicate a role for hepatic expression levels of one or more *Fmo* genes in both enhanced atherosclerotic plaque and decreased HDL levels in mice, the participation of *FMO* genes in human atherosclerosis and HDL cholesterol levels remains to be established. Strong associations between systemic TMAO levels and both angiographic measures of coronary artery atherosclerotic burden and cardiac risks were observed among subjects; however, no correlation was observed between plasma TMAO levels and HDL cholesterol levels in subjects. It remains to be determined whether genetic impairment in *FMO3* alone or in

©2011 Macmillan Publishers Limited. All rights reserved

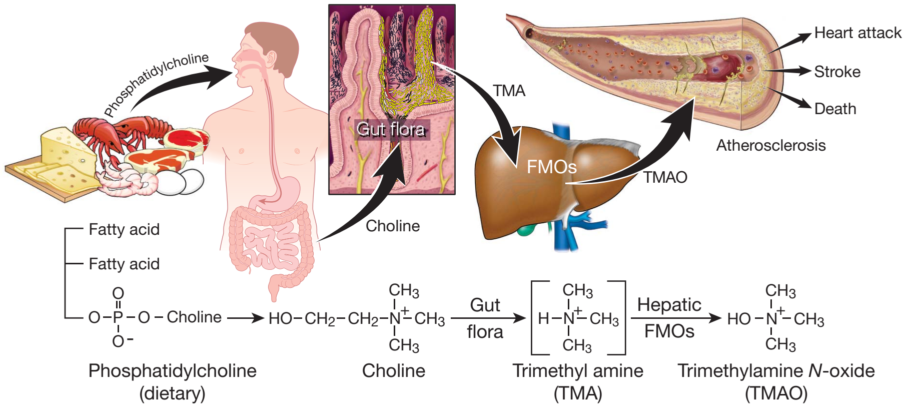


Figure 6 | Gut-flora-dependent metabolism of dietary PC and atherosclerosis. Schematic summary illustrating newly discovered pathway for gut-flora-mediated generation of pro-atherosclerotic metabolite from dietary PC.

combination with other *FMO* genes is protective for CVD. No phenotype other than the objectionable odour accompanying this disorder is known. In fact, individuals with trimethylaminuria often become vegans, as reducing the ingestion of dietary animal products rich in lipids decreases TMA production and the associated noxious odour. Little is also known about the biological functions of TMAO in humans. TMAO apparently serves as an osmolite in the freeze-avoidance response of some species[21]. *In vitro* it can function as a small-molecule chaperone, affecting the folding and functioning of some proteins[22,23]. In addition, TMAO and TMA accumulate in plasma of subjects on maintenance haemodialysis[24], raising the possibility that TMAO may contribute to the well-established enhanced CVD risk noted among subjects with end-stage renal disease.

Choline is an essential nutrient that is usually grouped within the vitamin B complex. Choline and its metabolite betaine are methyl donors, along with folate, and are metabolically linked to transmethylation pathways including synthesis of the CVD risk factor homocysteine. Deficiency in both choline and betaine have been suggested to produce epigenetic changes in genes linked to atherosclerosis[25,26], and acute choline and methionine deficiency in rodent models causes lipid accumulation in liver (steatohepatitis), heart and arterial tissues[27]. Alternatively, some studies have reported an association between increased whole blood levels of total choline and cardiovascular disease[28,29]. Few clinical studies have examined the relationship between choline intake and CVD[30], probably because accurate measures of the choline content of most foods has only recently become available[14] (http://www.nal.usda.gov/fnic/foodcomp/Data/Choline/Choln02.pdf). The association between dietary choline (and alternative trimethyl-amine-containing species) and atherosclerosis will be complex because, as the present studies show, it will be influenced by inter-individual differences in the composition of the intestinal microflora.

The human intestinal microbial community is an enormous and diverse ecosystem with known functions in nutrition, gut epithelial cell health, and innate immunity[31]. Intestinal flora recently also has been implicated in the development of some metabolic phenotypes such as obesity and insulin resistance, as well as alterations in immune responses[11,32–34]. To our knowledge, the present studies are the first to identify a direct link between intestinal microflora, dietary PC and CVD risk. These results indicate that an appropriately designed probiotic intervention may serve as a therapeutic strategy for CVD. Interestingly, production of TMAO can be altered by probiotic administration[35]. Thus, in addition to the current clinical recommendation for a general reduction in dietary lipids, manipulation of commensal microbial composition may be a novel therapeutic approach for the prevention and treatment of atherosclerotic heart disease and its complications. The present studies also suggest a further novel treatment for atherosclerosis—blocking the presumed pathogenic biochemical pathway at the level of the gut flora through use of a non-systemically absorbed inhibitor.

METHODS SUMMARY

Plasma samples and associated clinical study data were identified in patients referred for cardiac evaluation at a tertiary care centre. All subjects gave written informed consent and the Institutional Review Board of the Cleveland Clinic approved all study protocols. Unbiased metabolic profiling was performed using LC/MS. Target analyte structural identification was achieved using a combination of LC/MS/MS, LC/MS^n, multinuclear NMR, GC/MS and choline isotope tracer feeding studies in mice as outlined in Methods. Statistical analyses were performed using R (version 2.10.1)[36]. Intestinal microflora were suppressed by supplementation of drinking water with a cocktail of broad-spectrum antibiotics[37]. Germ-free mice were purchased from Taconic SWGF. QTL analyses to identify atherosclerosis-related genes were performed on F2 mice generated by crossing atherosclerosis-prone C57BL/6J *Apoe*$^{-/-}$ mice and atherosclerosis-resistant C3H/HeJ *Apoe*$^{-/-}$ mice[38]. mRNA expression was assayed by microarray analysis and real-time PCR. Aortic root lesion area in mice was quantified by microscopy after staining[39]. Mouse peritoneal macrophages were collected by lavage for foam cell quantification and cholesterol accumulation assay. Surface protein levels of scavenger receptors CD36 and SR-A1 were determined by flow cytometry.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 July 2009; accepted 9 February 2011.

1. Epstein, S. E. *et al.* The role of infection in restenosis and atherosclerosis: focus on cytomegalovirus. *Lancet* **348** (suppl. 1), S13–S17 (1996).
2. Patel, P. *et al.* Association of *Helicobacter pylori* and *Chlamydia pneumoniae* infections with coronary heart disease and cardiovascular risk factors. *Br. Med. J.* **311,** 711–714 (1995).
3. Danesh, J., Collins, R. & Peto, R. Chronic infections and coronary heart disease: is there a link? *Lancet* **350,** 430–436 (1997).
4. Saikku, P. *et al.* Serological evidence of an association of a novel *Chlamydia*, TWAR, with chronic coronary heart disease and acute myocardial infarction. *Lancet* **332,** 983–986 (1988).
5. O'Connor, C. M. *et al.* Azithromycin for the secondary prevention of coronary heart disease events—the WIZARD study: a randomized controlled trial. *J. Am. Med. Assoc.* **290,** 1459–1466 (2003).
6. Cannon, C. P. *et al.* Antibiotic treatment of *Chlamydia pneumoniae* after acute coronary syndrome. *N. Engl. J. Med.* **352,** 1646–1654 (2005).
7. Andraws, R., Berger, J. S. & Brown, D. L. Effects of antibiotic therapy on outcomes of patients with coronary artery disease: a meta-analysis of randomized controlled trials. *J. Am. Med. Assoc.* **293,** 2641–2647 (2005).
8. Wright, S. D. *et al.* Infectious agents are not necessary for murine atherogenesis. *J. Exp. Med.* **191,** 1437–1442 (2000).
9. Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science* **307,** 1915–1920 (2005).
10. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
11. Dumas, M. E. *et al.* Metabolic profiling reveals a contribution of gut microbiota to fatty liver phenotype in insulin-resistant mice. *Proc. Natl Acad. Sci. USA* **103,** 12511–12516 (2006).
12. Cashman, J. R. *et al.* Biochemical and clinical aspects of the human flavin-containing monooxygenase form 3 (FMO3) related to trimethylaminuria. *Curr. Drug Metab.* **4,** 151–170 (2003).
13. Al-Waiz, M., Mikov, M., Mitchell, S. C. & Smith, R. L. The exogenous origin of trimethylamine in the mouse. *Metabolism* **41,** 135–136 (1992).
14. Zeisel, S. H., Mar, M. H., Howe, J. C. & Holden, J. M. Concentrations of choline-containing compounds and betaine in common foods. *J. Nutr.* **133,** 1302–1307 (2003).

©2011 Macmillan Publishers Limited. All rights reserved

15. Lang, D. H. *et al.* Isoform specificity of trimethylamine *N*-oxygenation by human flavin-containing monooxygenase (FMO) and P450 enzymes: selective catalysis by fmo3. *Biochem. Pharmacol.* **56,** 1005–1012 (1998).
16. Zhang, A. Q., Mitchell, S. C. & Smith, R. L. Dietary precursors of trimethylamine in man: a pilot study. *Food Chem. Toxicol.* **37,** 515–520 (1999).
17. Mitchell, S. C. & Smith, R. L. Trimethylaminuria: the fish malodor syndrome. *Drug Metab. Dispos.* **29,** 517–521 (2001).
18. Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genet.* **37,** 710–717 (2005).
19. Dolphin, C. T., Janmohamed, A., Smith, R. L., Shephard, E. A. & Phillips, I. R. Missense mutation in flavin-containing mono-oxygenase 3 gene, *FMO3*, underlies fish-odour syndrome. *Nature Genet.* **17,** 491–494 (1997).
20. Wang, S. S. *et al.* Identification of pathways for atherosclerosis in mice: integration of quantitative trait locus analysis and global gene expression data. *Circ. Res.* **101,** e11–e30 (2007).
21. Treberg, J. R., Wilson, C. E., Richards, R. C., Ewart, K. V. & Driedzic, W. R. The freeze-avoidance response of smelt *Osmerus mordax*: initiation and subsequent suppression of glycerol, trimethylamine oxide and urea accumulation. *J. Exp. Biol.* **205,** 1419–1427 (2002).
22. Devlin, G. L., Parfrey, H., Tew, D. J., Lomas, D. A. & Bottomley, S. P. Prevention of polymerization of M and Z α_1-Antitrypsin (α_1-AT) with trimethylamine *N*-oxide. Implications for the treatment of α_1-AT deficiency. *Am. J. Respir. Cell Mol. Biol.* **24,** 727–732 (2001).
23. Song, J. L. & Chuang, D. T. Natural osmolyte trimethylamine *N*-oxide corrects assembly defects of mutant branched-chain α-ketoacid decarboxylase in maple syrup urine disease. *J. Biol. Chem.* **276,** 40241–40246 (2001).
24. Bain, M. A., Faull, R., Fornasini, G., Milne, R. W. & Evans, A. M. Accumulation of trimethylamine and trimethylamine-*N*-oxide in end-stage renal disease patients undergoing haemodialysis. *Nephrol. Dial. Transplant.* **21,** 1300–1304 (2006).
25. Dong, C., Yoon, W. & Goldschmidt-Clermont, P. J. DNA methylation and atherosclerosis. *J. Nutr.* **132,** 2406S–2409S (2002).
26. Zaina, S., Lindholm, M. W. & Lund, G. Nutrition and aberrant DNA methylation patterns in atherosclerosis: more than just hyperhomocysteinemia? *J. Nutr.* **135,** 5–8 (2005).
27. Salmon, W. D. & Newberne, P. M. Cardiovascular disease in choline-deficient rats. Effects of choline deficiency, nature and level of dietary lipids and proteins, and duration of feeding on plasma and liver lipid values and cardiovascular lesions. *Arch. Pathol.* **73,** 190–209 (1962).
28. Danne, O., Lueders, C., Storm, C., Frei, U. & Mockel, M. Whole blood choline and plasma choline in acute coronary syndromes: prognostic and pathophysiological implications. *Clin. Chim. Acta* **383,** 103–109 (2007).
29. LeLeiko, R. M. *et al.* Usefulness of elevations in serum choline and free F_2-isoprostane to predict 30-day cardiovascular outcomes in patients with acute coronary syndrome. *Am. J. Cardiol.* **104,** 638–643 (2009).
30. Bidulescu, A., Chambless, L. E., Siega-Riz, A. M., Zeisel, S. H. & Heiss, G. Usual choline and betaine dietary intake and incident coronary heart disease: the Atherosclerosis Risk in Communities (ARIC) study. *BMC Cardiovasc. Disord.* **7,** 20 (2007).
31. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308,** 1635–1638 (2005).
32. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444,** 1022–1023 (2006).
33. Li, M. *et al.* Symbiotic gut microbes modulate human metabolic phenotypes. *Proc. Natl Acad. Sci. USA* **105,** 2117–2122 (2008).
34. Reigstad, C. S., Lunden, G. O., Felin, J. & Backhed, F. Regulation of serum amyloid A3 (SAA3) in mouse colonic epithelium and adipose tissue by the intestinal microbiota. *PLoS ONE* **4,** e5842 (2009).
35. Martin, F. P. *et al.* Probiotic modulation of symbiotic gut microbial–host metabolic interactions in a humanized microbiome mouse model. *Mol. Syst. Biol.* **4,** 157 (2008).
36. Rizzo, M. L. *Statistical Computing with R* (Chapman & Hall/CRC, 2008).
37. Rakoff-Nahoum, S., Paglino, J., Eslami-Varzaneh, F., Edberg, S. & Medzhitov, R. Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. *Cell* **118,** 229–241 (2004).
38. Wang, S. *et al.* Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genet.* **2,** e15 (2006).
39. Baglione, J. & Smith, J. D. Quantitative assay for mouse atherosclerosis in the aortic root. *Methods Mol. Med.* **129,** 83–95 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We wish to thank E. Sehayek for discussions, L. W. Castellani for help with lipoprotein profile analysis, and F. McNally, M. Berk and M. Pepoy for technical assistance. This research was supported by National Institutes of Health grants R01 HL103866, P01 HL098055, P01HL087018-020001, P01 HL28481 and P01 HL30568. B.J.B. was supported by NIH training grant T32-DK07789. The clinical study GeneBank was supported in part by P01 HL076491-055328, R01 HL103931 and the Cleveland Clinic Foundation General Clinical Research Center of the Cleveland Clinic/Case Western Reserve University CTSA (1UL1RR024989). Some of the laboratory studies (haemaglobin A1C, fasting glucose) in GeneBank were supported by R01 DK080732 and Abbott Diagnostics provided supplies for performance of some of the fasting lipid profile, glucose, creatinine, troponin I and hsCRP measured in GeneBank.

Author Contributions Z.W. performed metabolomics analyses, and biochemical, cellular, animal model and mass spectrometry studies. He assisted with statistical analyses, and assisted in both drafting and critical review of the manuscript. E.K., B.D. and J.D.S. assisted with performance of animal models and their analyses. B.S.L. synthesized d9-DPPC and assisted in metabolomics/mass spectrometry analyses. B.J.B., H.A. and A.J.L. performed the mouse eQTL experiments and analyses, and assisted in both drafting and critical review of the manuscript. A.J.L. provided some funding for the study. R.K., E.B.B., X.F. and Y.-M.C. performed mass spectrometry analyses of clinical samples. Y.W. performed statistical analysis. A.E.F. and P.S. helped with collection of human liver biopsy material and interpretation of biochemical and pathological examination of animal liver for steatosis. W.H.W.T. assisted in GeneBank study design and enrolment, as well as analyses of clinical studies and critical review of the manuscript. J.A.D. assisted in clinical laboratory testing for human clinical studies, animal model experimental design, and critical review of the manuscript. S.L.H. conceived of the idea, designed experiments, assisted in data analyses, the drafting and critical review of the manuscript, and provided funding for the study.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.L.H. (hazens@ccf.org).

©2011 Macmillan Publishers Limited. All rights reserved

METHODS

General procedures. Lipids were extracted by chloroform:methanol (2:1, v/v)[40]. Cholesterol was quantified by GC/MS[41]. Triglyceride was quantified by GPO reagent set (Pointe Scientific)[42]. Cell DNA content was quantified by PicoGreen[43]. RNA was isolated by TRIZOL reagent (Invitrogen) and RNeasy Mini Kit (Qiagen). All reagents were purchased from Sigma unless otherwise specified.

Research subjects. Plasma samples and associated clinical data were collected as part of two studies involving stable non-symptomatic subjects undergoing elective cardiac evaluations at a tertiary care centre. The first study, GeneBank, is a large well-characterized tissue repository with longitudinal data from subjects undergoing elective diagnostic left heart catheterization or elective coronary computed tomography angiography[44]. The second study, BioBank, includes subjects undergoing cardiac risk factor evaluation/modification in a preventive cardiology clinic[45]. CAD included adjudicated diagnoses of stable or unstable angina, myocardial infarction or angiographic evidence of ≥50% stenosis of one or more epicardial vessels. PAD was defined as any evidence of extra-coronary atherosclerosis. Atherosclerotic CVD was defined as the presence of either CAD or PAD. All subjects gave written informed consent and the Institutional Review Board of the Cleveland Clinic approved all study protocols.

Discovery metabolomics analyses began with an unbiased search for plasma (fasting, EDTA purple top tube) analytes linked to CVD risk using a case–control design (Learning Cohort, $N = 50$ cases and 50 controls). Cases were randomly selected from GeneBank subjects who experienced a myocardial infarction, stroke or death over the ensuing 3-year period. An age- and gender-matched control group was randomly selected from GeneBank subjects that did not experience a CVD event. An independent non-overlapping Validation Cohort ($N = 25$ cases and 25 controls) was also from GeneBank. A third large ($N = 1{,}876$) independent study comprised of non-overlapping subjects then evaluated clinical associations of identified analytes. Approximately half ($N = 1{,}020$) of the subjects enrolled were from GeneBank and the remaining ($N = 856$) were from BioBank. Similar patient characteristics within each cohort and the combined cohort are observed, as shown in Supplementary Table 4a, b. The association of each PC metabolite and various cardiovascular phenotypes within each cohort (GeneBank and BioBank) are also similar (Supplementary Tables 4c–e). All subjects in the large independent clinical study had similar inclusion and exclusion criterion, negative cardiac enzymes (troponin I $< 0.03\ \text{ng ml}^{-1}$) and no recent history of myocardial infarction or coronary artery bypass graft. Estimate of glomerular filtration rate was calculated using the MDRD formula[46]. Fasting blood glucose, C reactive protein, troponin I and lipid profiles were measured on the Abbott ARCHITECT platform (Abbott Diagnostics).

Metabolomics analyses. Plasma proteins were precipitated with four volumes of ice-cold methanol and small-molecule analytes within supernatants were analysed after injection onto a phenyl column (4.6 × 250 mm, 5 μm Rexchrom Phenyl; Regis) at a flow rate of $0.8\ \text{ml min}^{-1}$ using a Cohesive HPLC interfaced with a PE Sciex API 365 triple quadrupole mass spectrometer (Applied Biosystems) with Ionics HSID+, EP10+, XT+ redesigned source and collision cell as upgrades in positive MS1 mode. LC gradient (LC1) starting from 10 mM ammonium formate over 0.5 min, then to 5 mM ammonium formate, 25% methanol and 0.1% formic acid over 3 min, held for 8 min, followed by 100% methanol and water washing for 3 min at a flow rate of $0.8\ \text{ml min}^{-1}$ was used to resolve analytes. Spectra were continuously acquired after the initial 4 min. Peaks within reconstructed ion chromatograms at 1 AMU increments were integrated and both retention times and m/z of analytes were used for statistical analyses.

Selection criteria for determining analytes of interest were based on the composite of MACE as the clinical phenotype, defined as incident myocardial infarction, stroke or death at 3 years, and included: (1) demonstration of a statistically significant difference between cases versus controls using a Bonferroni adjusted two sided *t*-test ($P < 0.05$); (2) evidence of a significant ($P < 0.05$) dose–response relationship between analyte level and clinical phenotype using Cochran–Armitage trend test; and (3) a minimal signal-to-noise ratio of 5:1 for a given analyte.

Chemical characterization of unknown metabolites. To chemically define the structures of the plasma analytes selected for further investigation (that is, analytes with m/z 76, 104 and 118 in positive MS1 mode), multiple approaches were used. Analytes of interest were isolated by HPLC, vacuum dried, re-dissolved in water and injected onto the same phenyl column with a distinct HPLC gradient (LC2, flow rate: $0.8\ \text{ml min}^{-1}$) starting from 0.2% formic acid over 2 min, then linearly to 18% acetonitrile containing 0.2% formic acid over 18 min and further to 100% acetonitrile containing 0.2% formic acid over 3 min. The targeted analytes were identified by their m/z and the appropriate fractions recovered. After removal of solvent, dry analytes were used for structural identification.

Samples analysed by GC/MS were derivatized using Sylon HTP kit (HMDS + TMCS + Pyridine (3: 1: 9), Supelco). Derivatization of TMAO and the plasma analyte at m/z 76 also included initial reduction by titanium (III) chloride[47] and further reaction with 2,2,2-trichloroethylchloroformate[48]. Analyses were performed on the Agilent Technolgies 6890/5973 GC/MS in positive ion chemical ionization mode. The GC/MS analyses used a J&W Scientific DB-1 column (30 m, 0.25-mm inner diameter, 0.25-μm film thickness) for separations.

Quantification of TMAO, choline and betaine. Stable isotope dilution LC/MS/MS was used for quantification of TMAO, choline and betaine. TMAO, choline and betaine were monitored in positive MRM MS mode using characteristic precursor–product ion transitions: m/z 76 → 58, m/z 104 → 60 and m/z 118 → 59, respectively. The internal standards TMAO-trimethyl-d_9 (d9-TMAO) and choline-trimethyl-d_9 (d9-choline), were added to plasma samples before protein precipitation, and were similarly monitored in MRM mode at m/z 85 → 68 and m/z 113 → 69, respectively. Various concentrations of TMAO, choline and betaine standards and a fixed amount of internal standards were spiked into control plasma to prepare the calibration curves for quantification of plasma analytes. TMA was similarly quantified from acidified plasma by LC/MS/MS using MRM mode.

Aortic root lesion quantification. Apolipoprotein E knockout mice (C57BL/6J *Apoe*−/−) were weaned at 4 weeks of age and fed with either standard chow control diet (Teklad 2018) or a custom diet comprised of normal chow supplemented with 0.5% choline (Teklad TD.07863), 1.0% choline (Teklad TD.07864) or 0.12% TMAO (Teklad TD.07865) for 16 weeks. Mice were anaesthetized with ketamine/xylazine before cardiac puncture to collect blood. Hearts were fixed and stored in 4% paraformaldehyde before frozen OCT sectioning and staining with Oil red O and haematoxylin. Aortic root lesion area was quantified as the mean value of six sections[39]. Aortic sections were immunostained with rat anti-mouse F4/80 antibody (ab6640, Abcam) followed by goat anti-rat IgG-FITC antibody (sc-2011, Santa Cruz) and FITC-conjugated CD36 monoclonal antibody (Cayman Chemical) for F4/80 and CD36, respectively. Sections were mounted in Vectashield DAPI (H-1200, Vectashield) to take pictures under a Leica DMR microscope (W. Nuhsbaum) equipped with a Q Imaging Retiga EX camera. We used Image-Pro Plus Version 7.0 (MediaCybernetics) to integrate the positive staining area of F4/80 and CD36 in aorta.

Flow cytometry assays on scavenger receptors. Cell surface expression of scavenger receptors SR-A1 and CD36 were quantified on peritoneal macrophages from female mice by flow cytometry after immunostaining with fluorochrome-conjugated antibodies. Fluorescence intensity was quantified on a FACSCalibur flow cytometry instrument with FlowJo software (BD Biosciences). More than 10,000 total events were acquired to obtain adequate macrophages numbers. The following antibodies were used to stain macrophages: CD36 monoclonal antibody FITC (Cayman Chemical), anti-mouse SR-AI/MSRA1 (R&D Systems), goat anti-rat IgG-FITC (Santa Cruz Biotechnology), Alexa Fluor 647 anti-mouse F4/80 (eBioscience), Alexa Fluor 647 anti-mouse CD11b (eBioscience) and the isotype controls, Alexa Fluor 647 rat IgG2b (eBioscience), Alexa Fluor 647 rat IgG2a (eBioscience), normal mouse IgA-FITC (Santa Cruz). Cells were incubated with antibodies for 30 min at 4 °C and washed with 0.1% BSA in PBS. Cells with double-positive staining for F4/80 and CD11b were gated as macrophage[49–51] for the quantification of fluorescence intensity for CD36 and SR-A1 (Supplementary Fig. 20), with results normalized to F4/80.

eQTL studies. C57BL/6J $Apoe^{-/-}$ (B6 $Apoe^{-/-}$) mice were purchased from the Jackson Laboratory and C3H/HeJ $Apoe^{-/-}$ (C3H $Apoe^{-/-}$) mice were bred by backcrossing B6 $Apoe^{-/-}$ to C3H/HeJ for 10 generations. F2 mice were generated by crossing B6 $Apoe^{-/-}$ with C3H $Apoe^{-/-}$ and subsequently intercrossing the F1 mice. Mice were fed Purina Chow containing 4% fat until 8 weeks of age, and then transferred to a Western diet (Teklad 88137) containing 42% fat and 0.15% cholesterol for 16 weeks until euthanasia at 24 weeks of age. Mouse atherosclerotic lesion area was quantified using standard methods[39]. eQTL analyses were performed as previously described[38]. Each individual sample was hybridized against the pool of F2 samples. Significantly differentially expressed genes were determined as previously described[52]. Expression data in the form of mean log ratios (mlratios) were treated as a quantitative trait in eQTL analysis using Rqtl package for the R language and environment for statistical computing (http://cran.r-project.org/).

Germ-free mice and conventionalization studies. An antibiotic cocktail ($0.5\ \text{g l}^{-1}$ vancomycin, $1\ \text{g l}^{-1}$ neomycin sulphate, $1\ \text{g l}^{-1}$ metronidazole, $1\ \text{g l}^{-1}$ ampicillin) previously shown to be sufficient to deplete all detectable commensal bacteria[37] was administered in drinking water ad libitum. In additional studies, 8-week-old female Swiss Webster germ-free mice (Taconic SWGF) underwent an oral (gavage) choline challenge (see later) immediately after their removal from their germ-free microisolator shipper. After the choline or PC challenge, the germ-free mice were placed in conventional cages with non-sterile C57BL/6J female mice to facilitate transfer of commensal organisms. Four weeks later, the conventionalized mice underwent a second choline or PC challenge.

©2011 Macmillan Publishers Limited. All rights reserved

***In vivo* macrophage studies.** C57BL/6J mice or B6 *Apoe* $^{-/-}$ mice were fed with either standard chow control diet (Teklad 2018) or a custom diet supplemented with 1.0% betaine (Teklad TD.08112), 1.0% choline (Teklad TD.07864), 0.12% TMAO (Teklad TD.07865) or 1.0% dimethylbutanol (DMB) supplemented in drinking water for at least 3 weeks. Elicited mouse peritoneal macrophages (MPMs) were harvested by peritoneal lavage with ice-cold PBS 3 days after i.p. injection of 1 ml 4% thioglycollate. Some studies with mice were performed using a custom diet with low but sufficient choline content (0.07% total; Teklad TD.09040) versus high-choline diet (1.0% total; Teklad TD.09041) in the presence or absence of antibiotics. Choline content of all diets was confirmed by LC/MS/MS.

Foam cell staining. Foam cells were identified by microscopy cultured peritoneal macrophages on glass coverslips after 6 h in RPMI 1640 medium supplemented with 3% lipoprotein-deficient serum. Cells were fixed with paraformaldehyde and stained with Oil red O/haematoxylin[53]. Cells containing $>$10 lipid droplets were scored as foam cells[50]. At least 10 fields and 500 cells per condition were counted.

Real-time PCR. Real-time PCR of *Cd36*, *Sr-a1* and flavin monooxygenases (mouse *Fmo*s) was performed using Brilliant II SYBR Green qRT–PCR kit (Strategene). The forward and reverse primers *Cd36*, *Gapdh*, *Sr-a1*, mouse *Fmo*s and *F4/80* were synthesized by IDT based on sequences reported[54–58]. RT–PCR of human FMOs was similarly performed using primers specific for the sequence of each of the indicated human FMOs.

d9-DPPC synthesis and vesicle preparation. d9-DPPC was synthesized by reacting 1,2-dipalmitoyl-sn-glycero-3-phosphoethanolamine (Genzyme Pharmaceuticals) with per-deuteromethyliodide (CD_3I, Cambridge Isotope Laboratories)[59,60]. The product was purified by preparative silica gel TLC and confirmed by both MS and NMR. Egg yolk lecithin (Avanti Polar Lipids) and d9-DPPC liposomes used for gavage feeding and i.p. injection of mice were prepared by the method of extrusion through polycarbonate filters[61].

Metabolic challenges in mice. C57BL/6J mice were administered (gavage) unlabelled or the indicated stable-isotope-labelled choline or PC (egg yolk lecithin or d9-DPPC) using a 1.5-inch 20-gauge intubation needle. Choline challenge: gavage consisted of 150 μl of 150 mM d9-choline. PC challenge: gavage or i.p. injection of 300 μl 5 mg ml^{-1} of unlabelled PC or labelled d9-DPPC. Mice were fasted 12 h before PC challenge. Plasma (50 μl) was collected via the saphenous vein from mice at baseline and after gavage or i.p. injection time points.

Statistical analysis. Student's *t*-test and Wilcoxon rank sum test were employed to compare group means[62,63]. Pearson correlation, Spearman rank correlation and Somers' Dxy correlation were used to investigate the correlation between two variables[64,65]. Comparison of categorical measures between independent groups was done using χ^2 tests[66]. Odds ratios and 95% confidence intervals for cardiovascular phenotypes (history of myocardial infarction, CAD, PAD, CVD and CAD+PAD) were calculated with R, version 2.10.1 (http://www.r-project.org), using logistic regression[67] with case status as the dependent variable and plasma analyte as independent variable. Trend tests in frequencies across quartiles were done using Cochran–Armitage trend tests[68]. Levels of analytes were adjusted for traditional CAD risk factors in a multivariate logistic regression model including individual traditional cardiac risk factors (age, gender, diabetes, smoking, hypertension, lipids, CRP and estimated creatinine clearance) and medication usage (statin or other lipid-lowering agents, antihypertensive agents including angiotensin-converting-enzyme inhibitor, angiotensin-receptor blocking agent, diuretic, calcium-channel blocker or beta blocker, and aspirin or other platelet inhibitors).

40. Folch, J., Lees, M. & Sloane Stanley, G. H. A simple method for the isolation and purification of total lipides from animal tissues. *J. Biol. Chem.* **226,** 497–509 (1957).
41. Robinet, P., Wang, Z., Hazen, S. L. & Smith, J. D. A simple and sensitive enzymatic method for cholesterol quantification in macrophages and foam cells. *J. Lipid Res.* **51,** 3364–3369 (2010).
42. Millward, C. A. *et al.* Genetic factors for resistance to diet-induced obesity and associated metabolic traits on mouse chromosome 17. *Mamm. Genome* **20,** 71–82 (2009).
43. Ahn, S. J., Costa, J. & Emanuel, J. R. PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. *Nucleic Acids Res.* **24,** 2623–2625 (1996).
44. Wang, Z. *et al.* Protein carbamylation links inflammation, smoking, uremia and atherogenesis. *Nature Med.* **13,** 1176–1184 (2007).
45. Nicholls, S. J. *et al.* Lipoprotein (a) levels and long-term cardiovascular risk in the contemporary era of statin therapy. *J. Lipid Res.* **51,** 3055–3061 (2010).
46. Stoves, J., Lindley, E. J., Barnfield, M. C., Burniston, M. T. & Newstead, C. G. MDRD equation estimates of glomerular filtration rate in potential living kidney donors and renal transplant recipients with impaired graft function. *Nephrol. Dial. Transplant.* **17,** 2036–2037 (2002).
47. Barham, A. H. *et al.* Appropriateness of cholesterol management in primary care by sex and level of cardiovascular risk. *Prev. Cardiol.* **12,** 95–101 (2009).
48. daCosta, K. A., Vrbanac, J. J. & Zeisel, S. H. The measurement of dimethylamine, trimethylamine, and trimethylamine *N*-oxide using capillary gas chromatography-mass spectrometry. *Anal. Biochem.* **187,** 234–239 (1990).
49. Schledzewski, K. *et al.* Lymphatic endothelium-specific hyaluronan receptor LYVE-1 is expressed by stabilin-1^{+}, F4/80^{+}, CD11b^{+} macrophages in malignant tumours and wound healing tissue *in vivo* and in bone marrow cultures *in vitro*: implications for the assessment of lymphangiogenesis. *J. Pathol.* **209,** 67–77 (2006).
50. Cailhier, J. F. *et al.* Conditional macrophage ablation demonstrates that resident macrophages initiate acute peritoneal inflammation. *J. Immunol.* **174,** 2336–2342 (2005).
51. Kunjathoor, V. V. *et al.* Scavenger receptors class A-I/II and CD36 are the principal receptors responsible for the uptake of modified low density lipoprotein leading to lipid loading in macrophages. *J. Biol. Chem.* **277,** 49982–49988 (2002).
52. Yang, X. *et al.* Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res.* **16,** 995–1004 (2006).
53. Zhou, J., Lhotak, S., Hilditch, B. A. & Austin, R. C. Activation of the unfolded protein response occurs at all stages of atherosclerotic lesion development in apolipoprotein E-deficient mice. *Circulation* **111,** 1814–1821 (2005).
54. Miles, E. A., Wallace, F. A. & Calder, P. C. Dietary fish oil reduces intercellular adhesion molecule 1 and scavenger receptor expression on murine macrophages. *Atherosclerosis* **152,** 43–50 (2000).
55. Westendorf, T., Graessler, J. & Kopprasch, S. Hypochlorite-oxidized low-density lipoprotein upregulates CD36 and PPARγ mRNA expression and modulates SR-BI gene expression in murine macrophages. *Mol. Cell. Biochem.* **277,** 143–152 (2005).
56. Rasooly, R., Kelley, D. S., Greg, J. & Mackey, B. E. Dietary *trans* 10, *cis* 12-conjugated linoleic acid reduces the expression of fatty acid oxidation and drug detoxification enzymes in mouse liver. *Br. J. Nutr.* **97,** 58–66 (2007).
57. Zhang, J. & Cashman, J. R. Quantitative analysis of *FMO* gene mRNA levels in human tissues. *Drug Metab. Dispos.* **34,** 19–26 (2006).
58. de Vries, T. J., Schoenmaker, T., Hooibrink, B., Leenen, P. J. & Everts, V. Myeloid blasts are the mouse bone marrow cells prone to differentiate into osteoclasts. *J. Leukoc. Biol.* **85,** 919–927 (2009).
59. Chen, F. C. M. & Benoiton, L. N. A new method of quatenizing amines and its use in amino acid and peptide chemistry. *Can. J. Chem.* **54,** 3310–3311 (1976).
60. Morano, C., Zhang, X. & Fricker, L. D. Multiple isotopic labels for quantitative mass spectrometry. *Anal. Chem.* **80,** 9298–9309 (2008).
61. Greenberg, M. E. *et al.* The lipid whisker model of the structure of oxidized cell membranes. *J. Biol. Chem.* **283,** 2385–2396 (2008).
62. Gauvreau, K. & Pagano, M. Student's *t*-test. *Nutrition* **9,** 386 (1993).
63. Wijnand, H. P. & van de Velde, R. Mann–Whitney/Wilcoxon's nonparametric cumulative probability distribution. *Comput. Methods Programs Biomed.* **63,** 21–28 (2000).
64. Gaddis, M. L. & Gaddis, G. M. Introduction to biostatistics: part 6, correlation and regression. *Ann. Emerg. Med.* **19,** 1462–1468 (1990).
65. Deichmann, M. *et al.* S100-β, melanoma-inhibiting activity, and lactate dehydrogenase discriminate progressive from nonprogressive American Joint Committee on Cancer stage IV melanoma. *J. Clin. Oncol.* **17,** 1891–1896 (1999).
66. Goodall, C. M., Stephens, O. B. & Moore, C. M. Comparative sensitivity of survival-adjusted chi-square and normal statistics for the mutagenesis fluctuation assay. *J. Appl. Toxicol.* **6,** 95–100 (1986).
67. Traissac, P., Martin-Prevel, Y., Delpeuch, F. & Maire, B. Logistic regression vs other generalized linear models to estimate prevalence rate ratios. *Rev. Epidemiol. Sante Publique* **47,** 593–604 (1999).
68. Gautam, S. Test for linear trend in 2 × K ordered tables with open-ended categories. *Biometrics* **53,** 1163–1169 (1997).

©2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09967

Streptococcal M1 protein constructs a pathological host fibrinogen network

Pauline Macheboeuf[1]†, Cosmo Buffalo[1], Chi-yu Fu[2], Annelies S. Zinkernagel[3]†*, Jason N. Cole[3,4]*, John E. Johnson[2], Victor Nizet[3,5] & Partho Ghosh[1]

M1 protein, a major virulence factor of the leading invasive strain of group A *Streptococcus*, is sufficient to induce toxic-shock-like vascular leakage and tissue injury. These events are triggered by the formation of a complex between M1 and fibrinogen that, unlike M1 or fibrinogen alone, leads to neutrophil activation. Here we provide a structural explanation for the pathological properties of the complex formed between streptococcal M1 and human fibrinogen. A conformationally dynamic coiled-coil dimer of M1 was found to organize four fibrinogen molecules into a specific cross-like pattern. This pattern supported the construction of a supramolecular network that was required for neutrophil activation but was distinct from a fibrin clot. Disruption of this network into other supramolecular assemblies was not tolerated. These results have bearing on the pathophysiology of streptococcal toxic shock.

The M protein[1] is the major surface-associated virulence factor of *Streptococcus pyogenes* (group A *Streptococcus*), a widespread bacterial pathogen that causes both mild infections and severe invasive diseases with high mortality rates (~30%), such as streptococcal toxic shock syndrome (STSS)[2]. Antigenic variation has resulted in >100 M protein types[3] but only a few are frequently associated with invasive disease, with the M1 type being the most prevalent[4]. Strains belonging to a globally disseminated subclone of the M1T1 serotype have been the leading cause of severe invasive group A *Streptococcus* infection worldwide for the past 30 years[5]. The M1 protein itself has pro-inflammatory properties and in animal models is sufficient to trigger vascular leakage and tissue injury similar to that observed in STSS[6–9]. These pathological properties of M1 require its interaction with fibrinogen. The M1–fibrinogen complex binds β_2 integrins on neutrophils and triggers the release of heparin binding protein (HBP), a potent vasodilator[10] and a strong indicator of sepsis and circulatory failure in patients[11]. The M1–fibrinogen complex also activates platelets in an integrin-dependent manner[12], leading to further activation of neutrophils as well as monocytes.

How the M1–fibrinogen complex causes neutrophil activation, when neither protein alone does[6,8], is not known. To address this issue, we determined the ~3.3 Å resolution crystal structure of an M1–fibrinogen complex containing the M1 fragment $M1^{BC1}$ (residues 132–263, ~17 kDa) and fibrinogen fragment D (FgD, ~86 kDa)[13,14] (Supplementary Table 1 and Supplementary Figs 1 and 2). The $M1^{BC1}$ fragment contains the B repeats, which are sufficient to bind fibrinogen[8,15], and the S region, to which immunoglobulin G (IgG) molecules bind and enhance the release of HBP through FcγRII[7]; it also contains the first C repeat of M1. FgD, which comprises the majority of fibrinogen, is necessary and sufficient to bind M1 (ref. 16). The final model consists of $M1^{BC1}$ residues 132–238, the register of which was separately determined using anomalous scattering (see Methods); all but the first five residues of the C repeats were apparently removed by proteolysis. The entirety of FgD, as seen in the crystal structure of the unbound form[13,14], was visible except for several residues at either end. Whereas residues of $M1^{BC1}$ distal to the interface with FgD were often in incomplete electron density, residues at the interface had well defined electron density, enabling specific intermolecular contacts to be discerned.

Cross-like pattern

The most striking aspect of the structure is the fact that $M1^{BC1}$ is surrounded by four FgD molecules in a cross-like pattern (Fig. 1). Two pairs of FgD molecules (each ~130 Å long) lie roughly perpendicular to one other as well as to $M1^{BC1}$ (~160 Å long), which runs through the centre of the cross. FgD, as previously described[13], consists of a parallel heterotrimeric (αβγ) α-helical coiled-coil connected to globular heads. $M1^{BC1}$ forms a parallel homodimeric α-helical coiled-coil throughout most of its length, including the B repeats that bind FgD. There are four B repeats in all, two per M1 chain, explaining the 2:4 $M1^{BC1}$:FgD stoichiometry of this ~380-kDa complex. The upstream B repeats (B1) bind two FgD molecules that are oriented ~180° to one another due to the dyad symmetry of the M1 coiled coil, as do the downstream B repeats (B2). B1 and B2 are separated by 28 residues, roughly a one-quarter turn of the coiled-coil superhelix, meaning that the two pairs of FgD molecules are oriented ~90° to one another, giving rise to the cross-like pattern.

The four separate M1–FgD contact sites are nearly identical in structure, being predominantly polar and each having a small buried surface area of ~645 Å^2 (Fig. 2a). Residues in the B repeats from both helices of the coiled coil and every heptad position (that is, *a*–*g*) contribute to FgD binding (Fig. 2b). Although the two B repeats are imperfect in sequence, the FgD-binding residues are identical between the two (Fig. 3). For FgD, the coiled coils in the β and γ chains are involved in binding M1, and contribute residues from the exposed *b*, *c* and *e* positions to the interface. The site at which M1 binds FgD is notably quite distant (~90 Å) from the fibrinogen γC globular head, which has been shown to bind β_2 integrins[17]. The M1 S

[1] Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, USA. [2] Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, California 92037, USA. [3] Department of Pediatrics, University of California, San Diego, La Jolla, California 92093, USA. [4] School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, Queensland 4072, Australia. [5] School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, California 92093, USA. †Present addresses: Unit of Virus Host Cell Interactions, UMI 3265, Université Joseph Fourier-EMBL-CNRS, 6 rue Jules Horowitz 38042 Grenoble, France (P.M.); University Hospital Zurich, University Zurich, Rämistrstr. 100, 8091 Zurich, Switzerland (A.S.Z.).

*These authors contributed equally to this work.

©2011 Macmillan Publishers Limited. All rights reserved

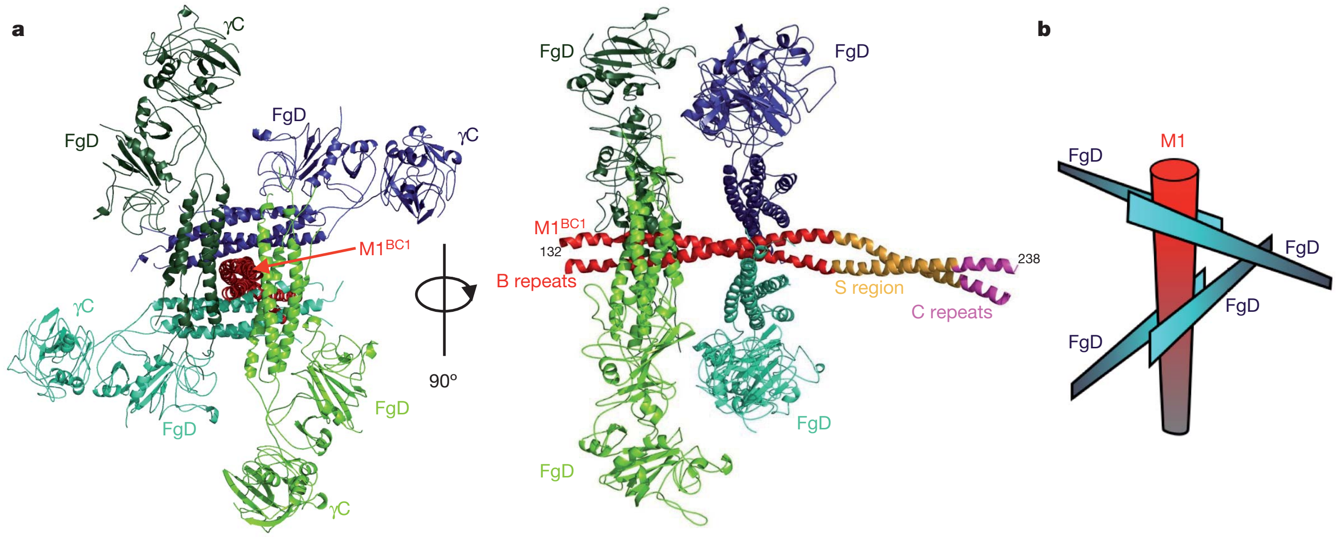


Figure 1 | M1 assembles fibrinogen into a cross-like pattern. **a**, The M1^{BC1}–FgD structure in ribbon representation. The M1^{BC1} B repeats are in red, S region in gold, and C repeats in purple. The four FgD molecules bound by M1^{BC1} are in shades of blue or green, and the β_2 integrin-binding γC domains are indicated. **b**, Schematic of the cross-like pattern of FgD (blue blades) surrounding M1^{BC1} (red cylinder).

region is freely available to bind IgG molecules and thereby enhance HBP release.

Conformational dynamics

The heptad register in the B repeats observed here differs from that observed previously in the crystal structure of the M1^{AB} fragment (ref. 8), which includes the A region and B repeats but not the S region and C repeats. Residues that were in the *d*–*g* heptad face of M1^{AB} (register 1) occupy the *a*–*d* face of M1^{BC1} bound to FgD (register 2) (Figs 3a, b). These two registers are related by a rotation of one helical face, or ~51.4° (Fig. 3c). The ability of the B repeats to adopt these two competing registers is supported by coiled-coil propensity analysis[18],

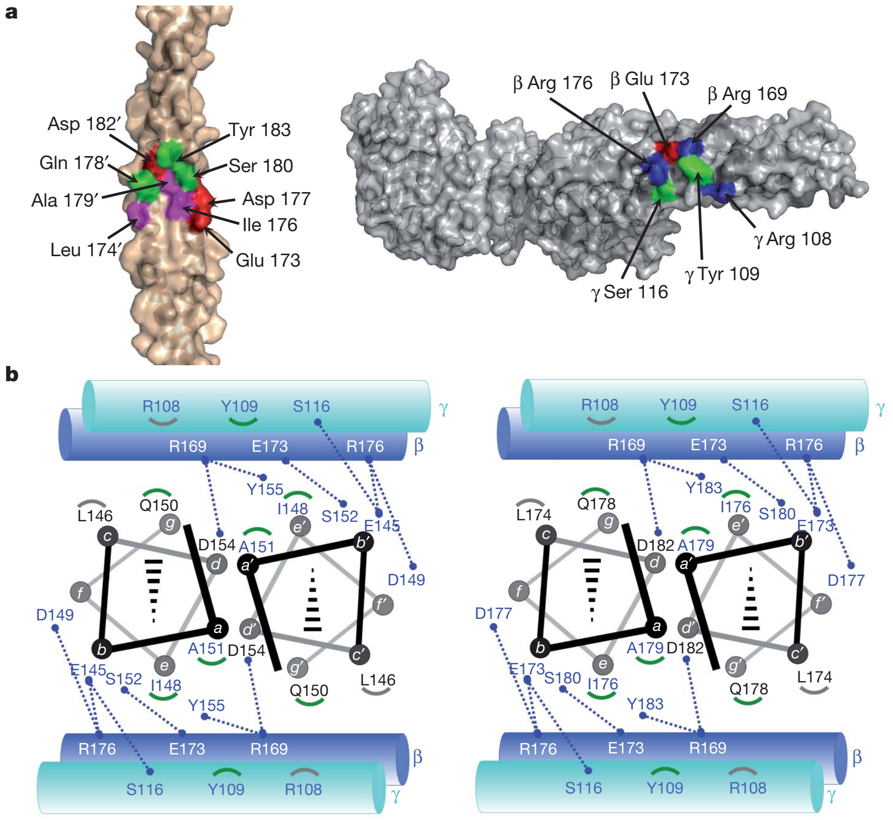


Figure 2 | M1–fibrinogen interface. **a**, Interface between M1^{BC1} B2 (left, primed numbers refer to one helix and non-primed the opposing helix) and FgD (right) in surface representation (basic residues, blue; acidic, red; polar, green; non-polar, magenta). **b**, Schematic of the interface between M1^{BC1} B1 (left) and B2 (right) in helical projection and FgD (cylinders, β-chain in blue and γ in blue–green). Blue dotted lines connect residues making polar contacts, and grey and green arcs correspond to M1 residues making van der Waals contacts to fibrinogen γ 108 and 109, respectively.

©2011 Macmillan Publishers Limited. All rights reserved

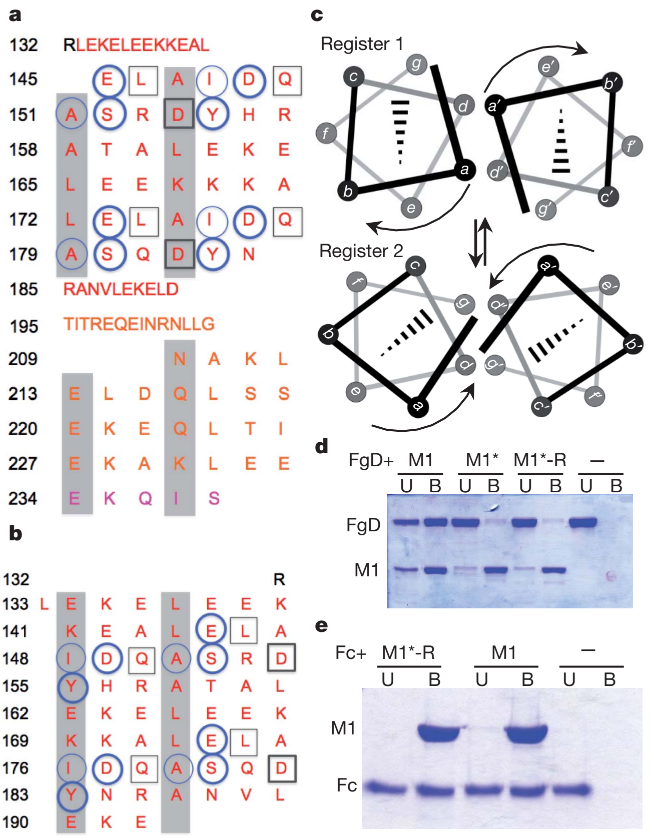


Figure 3 | Conformational dynamics. **a**, Heptad register of M1^{BC1} bound to FgD (register 2; *a* and *d* residues shaded grey). Residues not assigned a heptad position (132–144, 185–208) form an α-helical dimer but do not have coiled-coil 'knobs-into-holes' packing. FgD-contacting residues are in circles for one M1 helix and in boxes for the other. Heavy lines denote polar contacts, and light lines van der Waals contacts. The B repeats are in red, the S region in gold, and C repeat in purple. **b**, Heptad register of unbound M1^{AB} (register 1; ref. 8). **c**, Relationship between registers 1 and 2. **d**, Association of His-tagged M1, M1* and M1*-R with FgD at 37 °C, as assessed by a Ni^{2+}-nitrilotriacetic acid (NTA) agarose co-precipitation assay and visualized by non-reducing, Coomassie-stained SDS–PAGE. U, unbound fraction; B, bound fraction. **e**, Association of His-tagged M1 and M1*-R with IgG Fc at 37 °C, as in **d**, except visualized by reducing, Coomassie-stained SDS–PAGE.

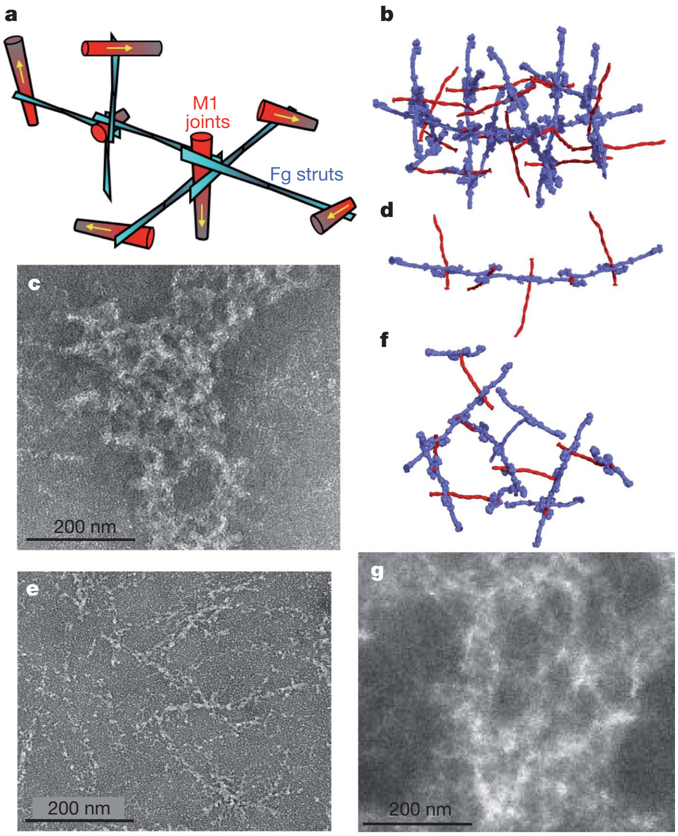


Figure 4 | M1–fibrinogen network. **a**, Schematic of the M1–fibrinogen network. The yellow arrows specify the direction of M1. **b**, **c**, Model of the M1–fibrinogen network, with M1 (red) and fibrinogen (blue) in surface representation (**b**), and negative-stained electron micrograph of M1 co-incubated with fibrinogen (**c**). **d**–**g**, Identical to **b**, **c**, except with ΔB2 (**d**, **e**) and B2C (**f**, **g**) modelled and co-incubated with fibrinogen.

which indicates that both registers 1 and 2 are embedded within the B repeats as short interspersed stretches (Supplementary Fig. 3). Surprisingly, residues that bind fibrinogen have a preference for register 1, which is incapable of binding fibrinogen, but are surrounded by residues that have a preference for register 2, the fibrinogen-binding register. In addition to these two registers being alternately sampled by the B repeats, a splayed conformation probably exists, as suggested by the dynamic dissociation and reassociation of M1 chains[8]. Presumably the splayed conformation enables transitions between registers 1 and 2, the latter being stabilized by fibrinogen binding.

To test experimentally for the presence of conformational dynamics, we stabilized register 1 in the B repeats without altering fibrinogen-binding residues. We hypothesized that fibrinogen binding should be decreased through this process if register 1 were sampled, because M1 would be 'locked' in the non-binding register. The ideal coiled-coil residues Val and Leu were substituted at *a* and *d* positions, respectively, of register 1 in the B repeats[19,20], except at the six *a* and *d* positions involved in FgD binding (Supplementary Fig. 4). Most notable among these six were Tyr 155 and Tyr 183, which are at core *a* positions in register 1 but at exposed *e* positions in register 2, from which they make π-cation interactions with FgD β169. This variant of M1, called M1*-revertant (M1*-R), is equivalent to the previously characterized M1* except with all the fibrinogen-binding residues present[8]. M1* was shown to be more stable than wild-type M1 but substantially diminished in FgD binding. We found that M1*-R, like M1*, was also greatly attenuated for FgD binding (Fig. 3d), and that this attenuation was specific, as M1*-R maintained wild-type levels of interaction with IgG Fc fragments (Fig. 3e). This latter interaction occurs through M1 protein regions outside the B repeats[8]. These mutational results support the conclusion that register 1, along with register 2, is sampled in the B repeats. Altogether our observations provide evidence for large conformational dynamics in the B repeats. What purpose these dynamics serve is unknown, but one possibility is that they are advantageous for group A *Streptococcus* immune evasion, in effect providing a 'moving target' for antibody recognition.

M1–fibrinogen network

To address the mechanism of neutrophil activation, we modelled fibrinogen and intact M1 in place of FgD and M1^{BC1} (Fig. 4 and Supplementary Movie 1), respectively. Importantly, because fibrinogen is a dimer (of αβγ heterotrimers), a fibrinogen molecule has two M1-binding sites as opposed to the single site in FgD. From this modelling emerged a non-clashing M1–fibrinogen network with fibrinogen acting as struts and M1 acting as joints. The two M1 molecules that bound an individual fibrinogen were tilted with respect to each other due to the inherent flexibility of fibrinogen[21]. This tilt gave the network three-dimensional character and meant that the network incorporated M1 molecules pointing in opposite directions. The variation in M1 directionality indicates that the network is formed by free M1 released from the bacterial surface by neutrophil proteases[6], as opposed to M1 anchored unidirectionally by its carboxy terminus to

©2011 Macmillan Publishers Limited. All rights reserved

the bacterial cell wall[22]. Consistent with this notion, the greater proportion of M1 in samples from STSS patients occurs as free released protein[7].

The structure of the M1–fibrinogen network indicated a mechanism for neutrophil activation. Previous work had shown that antibody crosslinking of β_2 integrin had the same effect on neutrophil activation as the M1–fibrinogen complex[6,9,23], indicating that β_2 integrin clustering and avidity are involved in signalling by M1–fibrinogen. On the basis of these data, we surmised that the fibrinogen density induced by the M1–fibrinogen network was likely to be a critical factor for neutrophil activation. To test this model, we compared HBP release by neutrophils stimulated with various M1 deletion constructs. M1 in which either the upstream or downstream B repeat was deleted, ΔB1 and ΔB2, respectively, retained fibrinogen binding due to the continued presence of one of the B repeats (Fig. 5a). However, as the modelling predicted, ΔB1 and ΔB2 formed fibres (Fig. 4d, e and Supplementary Fig. 5a) rather than the networks formed by wild-type M1 (Fig. 4b, c). Despite being able to bind fibrinogen, neither ΔB1 nor ΔB2 triggered release of HBP from neutrophils, which was in sharp contrast to wild-type M1 (Fig. 5b). This result indicates that the M1–fibrinogen network rather than fibrinogen-binding itself is required for neutrophil activation. We also demonstrated that the addition of FgD, which is unable to support network formation because it has only one M1-binding site, blocks M1-mediated neutrophil activation (Fig. 5c). Excess FgD was necessary in this experiment as binding of M1 to FgD is weaker than it is to fibrinogen[24], the difference being explained by the high avidity between M1 and fibrinogen as compared to the weaker affinity between M1 and FgD.

As expected, deletion of both B repeats (ΔB1B2) resulted in no networks, no fibres, and no induction of HBP release (Fig. 5b and Supplementary Fig. 5b). However, ΔB1B2 retained a low level of fibrinogen binding (Fig. 5a). On the basis of this and other evidence, we uncovered a cryptic fibrinogen-binding site in the A region. A molecular replacement solution of a low-resolution crystal (7.5 Å resolution limit) of the M1 A region bound to FgD (M1^{A}–FgD) confirmed the existence of this site. This solution revealed two molecules of FgD oriented 180° to one another and arranged perpendicularly to the A region, similar to the binding mode observed for each of the B repeats (Supplementary Fig. 6). Although the low resolution limited our abilities to discern A-region residues, it was apparent that the same FgD residues bound by the B repeats were bound by the A region. This indicated that A-region residues 106–119 are the likely fibrinogen-binding site (Supplementary Fig. 7), as this region has some sequence similarity to the B repeats, including a Tyr capable of forming a π-cation bond to FgD β169; tyrosines are otherwise rare in the M1 sequence. In line with observations for the B repeats, the A-region site would also require a ~51.4° rotation in helical register from the conformation observed in M1^{AB} (ref. 8) to bind fibrinogen. Deletion of this putative fibrinogen-binding site in the A region along with both B repeats completely abrogated fibrinogen binding (Fig. 5a, Δ98ΔB1B2). Although we found that the A-region site was not required for network formation or for HBP release (Fig. 5b, Δ98; see Supplementary Fig. 5c), the possibility that this cryptic site has other functions related to fibrinogen binding (for example, evasion of phagocytosis) merits future exploration.

Lastly, we asked whether the density of the M1–fibrinogen network was consequential. To address this, we deleted the downstream B repeat (B2) and inserted it at the C terminus of M1, thereby increasing the spacing between the two B repeats (Fig. 5a). Modelling predicted that a sparser network should be formed by this construct, called B2C. B2C bound FgD and formed networks, but notably, did not trigger release of HBP (Figs 4f, g and 5d). We note that the resolution of the electron micrographs did not allow us to distinguish between M1–fibrinogen and B2C–fibrinogen networks, but our modelling strongly suggests that the difference in network density accounted for the lack of neutrophil activation.

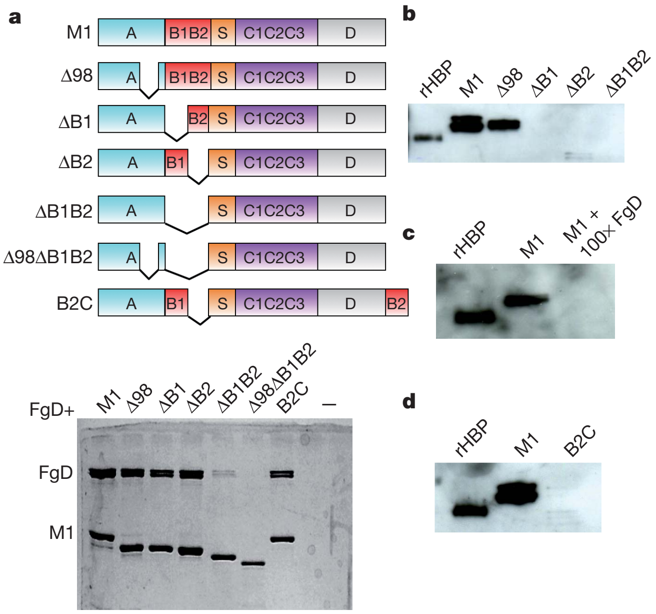


Figure 5 | **Fibrinogen binding and neutrophil activation.** **a**, Schematic of M1 constructs (top), with domains denoted. Bottom: association of His-tagged M1, Δ98 (Δ98–125), ΔB1 (Δ133–161), ΔB2 (Δ162–189), ΔB1B2 (Δ133–189), Δ98ΔB1B2 (Δ98–125, Δ133–189) and B2C (residues 162–189 deleted and inserted after C-terminal residue 453) with FgD as assessed by a Ni^{2+}-NTA agarose co-precipitation assay and visualized by non-reducing, Coomassie-stained SDS–PAGE. Only bound fractions are shown. **b**, Release of HBP by human neutrophils incubated with M1, Δ98, ΔB1, ΔB2, or ΔB1B2, as assayed by an anti-HBP western blot. The leftmost lane contains recombinant HBP (rHBP) as a positive control. The difference between this and other HBP samples is due to glycosylation. **c**, **d**, Release of HBP inhibited by a 100-fold excess of FgD (**c**), and elicited by B2C (**d**), both visualized as in **b**.

Conclusions

We have shown that a key pro-inflammatory property of M1, as exemplified by the induction of neutrophil HBP release, is due to the organization of fibrinogen into a specific cross-like pattern that supports the formation of an M1–fibrinogen network. This process requires the presence of two appropriately spaced B repeats. Repeats are a common feature of M protein sequences, but whether other fibrinogen-binding M protein types with repeat sequences have similar pro-inflammatory capabilities is unknown and under investigation. Because disruption of the M1–fibrinogen network into fibres or sparse networks resulted in loss of neutrophil activation, we conclude that the density of fibrinogen in the network is the critical factor in neutrophil activation. An alternative model would require a conformational change in fibrinogen upon M1 binding, as has been suggested for the unmasking of a β_2 integrin-binding tail in the γC globular domain of fibrinogen[25]. This possibility cannot be excluded as the tail is absent in FgD, but it seems unlikely to us as the tail would be quite distant from M1 and no large conformational changes were evident in FgD to transmit a binding signal to the tail. Although the M1–fibrinogen network is distinct from a fibrin clot, these supramolecular assemblies both present high densities of integrin-binding sites, indicating that integrin clustering and avidity are conserved mechanisms for leukocyte activation. Interference with the M1–fibrinogen interaction visualized here represents a potential therapeutic target to ameliorate the severe outcomes of STSS.

METHODS SUMMARY

Preparation of M1 and FgD, fibrinogen binding assays, and HBP release assays were carried out as previously described[8]. Diffraction data from crystals of M1^{BC1}–FgD and M1^{A}–FgD were processed using MOSFLM[26] or HKL2000[27],

©2011 Macmillan Publishers Limited. All rights reserved

and phases were determined by molecular replacement using the program Phaser[28] and the structure of human FgD[14] (Protein Data Bank code 3E1I) as a search model. The register of M1^{BC1} was verified by an anomalous dispersion experiment using crystals of selenomethionine-substituted M1^{BC1}(I148M)–FgD. Samples for electron microscopy were negatively stained with 0.2% uranyl acetate and imaged using a FEI Tecnai F20 Twin transmission electron microscope at an accelerating voltage of 120 kV.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 August 2010; accepted 22 February 2011.

1. Fischetti, V. A. Streptococcal M protein: molecular design and biological behavior. *Clin. Microbiol. Rev.* **2,** 285–314 (1989).
2. Cunningham, M. W. Pathogenesis of group A streptococcal infections. *Clin. Microbiol. Rev.* **13,** 470–511 (2000).
3. Facklam, R. F. *et al.* Extension of the Lancefield classification for group A streptococci by addition of 22 new M protein gene sequence types from clinical isolates: emm103 to emm124. *Clin. Infect. Dis.* **34,** 28–38 (2002).
4. Steer, A. C., Law, I., Matatolu, L., Beall, B. W. & Carapetis, J. R. Global emm type distribution of group A streptococci: systematic review and implications for vaccine development. *Lancet Infect. Dis.* **9,** 611–616 (2009).
5. Aziz, R. K. & Kotb, M. Rise and persistence of global M1T1 clone of *Streptococcus pyogenes*. *Emerg. Infect. Dis.* **14,** 1511–1517 (2008).
6. Herwald, H. *et al.* M protein, a classical bacterial virulence determinant, forms complexes with fibrinogen that induce vascular leakage. *Cell* **116,** 367–379 (2004).
7. Kahn, F. *et al.* Antibodies against a surface protein of *Streptococcus pyogenes* promote a pathological inflammatory response. *PLoS Pathog.* **4,** e1000149 (2008).
8. McNamara, C. *et al.* Coiled-coil irregularities and instabilities in group A *Streptococcus* M1 are required for virulence. *Science* **319,** 1405–1408 (2008).
9. Soehnlein, O. *et al.* Neutrophil degranulation mediates severe lung damage triggered by streptococcal M1 protein. *Eur. Respir. J.* **32,** 405–412 (2008).
10. Gautam, N. *et al.* Heparin-binding protein (HBP/CAP37): a missing link in neutrophil-evoked alteration of vascular permeability. *Nature Med.* **7,** 1123–1127 (2001).
11. Linder, A., Christensson, B., Herwald, H., Bjorck, L. & Akesson, P. Heparin-binding protein: an early marker of circulatory failure in sepsis. *Clin. Infect. Dis.* **49,** 1044–1050 (2009).
12. Shannon, O. *et al.* Severe streptococcal infection is associated with M protein-induced platelet activation and thrombus formation. *Mol. Microbiol.* **65,** 1147–1157 (2007).
13. Spraggon, G., Everse, S. J. & Doolittle, R. F. Crystal structures of fragment D from human fibrinogen and its crosslinked counterpart from fibrin. *Nature* **389,** 455–462 (1997).
14. Bowley, S. R. & Lord, S. T. Fibrinogen variant BßD432A has normal polymerization but does not bind knob ''B''. *Blood* **113,** 4425–4430 (2009).
15. Ringdahl, U. *et al.* A role for the fibrinogen-binding regions of streptococcal M proteins in phagocytosis resistance. *Mol. Microbiol.* **37,** 1318–1326 (2000).
16. Akesson, P., Schmidt, K. H., Cooney, J. & Bjorck, L. M1 protein and protein H: IgGFc- and albumin-binding streptococcal surface proteins encoded by adjacent genes. *Biochem. J.* **300,** 877–886 (1994).
17. Medved, L., Litvinovich, S., Ugarova, T., Matsuka, Y. & Ingham, K. Domain structure and functional activity of the recombinant human fibrinogen γ-module (γ148–411). *Biochemistry* **36,** 4685–4693 (1997).
18. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252,** 1162–1164 (1991).
19. Tripet, B., Wagschal, K., Lavigne, P., Mant, C. T. & Hodges, R. S. Effects of side-chain characteristics on stability and oligomerization state of a *de novo*-designed model coiled-coil: 20 amino acid substitutions in position ''d''. *J. Mol. Biol.* **300,** 377–402 (2000).
20. Wagschal, K., Tripet, B., Lavigne, P., Mant, C. & Hodges, R. S. The role of position a in determining the stability and oligomerization state of α-helical coiled coils: 20 amino acid stability coefficients in the hydrophobic core of proteins. *Protein Sci.* **8,** 2312–2329 (1999).
21. Kollman, J. M., Pandi, L., Sawaya, M. R., Riley, M. & Doolittle, R. F. Crystal structure of human fibrinogen. *Biochemistry* **48,** 3877–3886 (2009).
22. Navarre, W. W. & Schneewind, O. Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiol. Mol. Biol. Rev.* **63,** 174–229 (1999).
23. Gautam, N., Herwald, H., Hedqvist, P. & Lindbom, L. Signaling via β_2 integrins triggers neutrophil-dependent alteration in endothelial barrier function. *J. Exp. Med.* **191,** 1829–1840 (2000).
24. Whitnack, E. & Beachey, E. H. Inhibition of complement-mediated opsonization and phagocytosis of *Streptococcus pyogenes* by D fragments of fibrinogen and fibrin bound to cell surface M protein. *J. Exp. Med.* **162,** 1983–1997 (1985).
25. Lishko, V. K., Kudryk, B., Yakubenko, V. P., Yee, V. C. & Ugarova, T. P. Regulated unmasking of the cryptic binding site for integrin αMβ2 in the γC-domain of fibrinogen. *Biochemistry* **41,** 12942–12951 (2002).
26. Leslie, A. Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 ESF-EAMCB Newslett. Protein Crystallogr.* **26,** (1992).
27. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
28. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Cryst.* **40,** 658–674 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the Argonne Photon Source Laboratory GM/CA CAT staff for help with data collection, R. Doolittle and A. Royant for advice, and S. Mel and G. Ghosh for comments on the manuscript. This work was supported by NIH R21 AI071167 (P.G.), T32 GM007240 (C.B.), R01 AI077780 (V.N.), R01 GM54076 (J.E.J.), and a fellowship (J.N.C.) from the National Health and Medical Research Council of Australia (514639).

Author Contributions P.M. and P.G. designed the experiments. P.M. carried out the structure determination, modelling and FgD co-precipitation assay. C.B. carried out the Fc co-precipitation assay. P.M. and C.-y.F. carried out the electron microscopy under the supervision of J.E.J. P.M., A.S.Z. and J.N.C. carried out the HBP release assays under the supervision of V.N. P.M. and P.G. wrote the manuscript, and the other authors provided editorial advice.

Author Information Atomic coordinates and structure factors for M1^{BC1}–FgD (2XNX) and M1^{A}–FgD (2XNY) have been deposited with the Protein Data Bank. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to P.G. (pghosh@ucsd.edu).

©2011 Macmillan Publishers Limited. All rights reserved

METHODS

DNA manipulation. The DNA sequences of intact mature M1 protein (residues 42–453), M1^{BC1} (residues 128–263), and M1^{A} (residues 42–132) were cloned as described previously[8] from human group A *Streptococcus* isolate 5448 (ref. 29) into pET28b (Novagen). The M1 deletion mutants and B2C were generated using the QuickChange II Site-Directed mutagenesis kit (Stratagene), according to manufacturer's instructions, or by the mega-primer method[30]. All M1 protein constructs had a C-terminal His tag for purification purposes, except for M1^{BC1}.

Protein expression and purification. M1 protein constructs were expressed in *Escherichia coli* BL21 (DE3), which were grown in LB containing 34 mg ml^{-1} kanamycin at 37 °C until mid-logarithmic phase and then induced at room temperature with 1 mM isopropyl β-D-1-thiogalactopyranoside and grown further for 18 h. Bacteria were harvested by centrifugation and re-suspended in either 100 mM NaCl, 50 mM sodium phosphate buffer, pH 8 (SP) or 100 mM NaCl, 50 mM Tris, pH 8 (ST), both with protease inhibitors (Complete tablet, Roche), for biochemical analysis or crystallization experiments, respectively. Bacteria were lysed using an EmulsiFlex-C5 (Avestin). His-tagged M1 constructs were then purified as previously reported[8]. For M1^{BC1}, which lacks a His tag, the lysate was heat denatured at 75 °C for 30 min, cooled on ice for 30 min, and clarified by centrifugation. Nucleic acids were removed by the addition of 0.5% polyethyleneimine and the resulting supernatant was then precipitated with 75% $(NH_4)_2SO_4$. Precipitated protein was re-suspended in either SP or ST and dialysed overnight in the same buffer. Proteins were then purified on a Q-Sepharose anion exchange column (GE-Healthcare). Selenomethionine incorporation into M1^{BC1} was carried out as previously described[31], and the purification was carried out as above except with the addition of 1 mM dithiothreitol throughout.

Fibrinogen fragment D (FgD) was purified from human fibrinogen as described previously[32]. Briefly, human fibrinogen (Calbiochem) was trypsinized overnight in 150 mM NaCl, 5 mM $CaCl_2$, 50 mM imidazole, pH 7 and purified on a Gly-Pro-Arg column equilibrated with the same buffer. The protein was eluted from the column with 1 M NaBr, 0.05 M NaOAc, pH 5.3 and subsequently exchanged by ultrafiltration into ST.

Crystallization and data collection. For crystallization of M1^{BC1}–FgD, M1^{BC1} was mixed with FgD at a 4:1 molar ratio, and the FgD–M1^{BC1} complex was purified on a Superdex 200 16/60 size exclusion column (Amersham) in 20 mM NaCl, 10 mM Tris, pH 8. Crystallization was performed by the vapour-diffusion method in two steps. The first step involved mixing of an equal volume of M1^{BC1}–FgD at 8 mg ml^{-1} and precipitant solution containing 0.6 M K_2/Na_2HPO_4, 0.12 M $(NH_4)_2SO_4$, 0.1 M HEPES, pH 7.5. This produced crystals that did not diffract X-rays. These non-diffracting crystals were crushed and diluted in the precipitant solution before being used as a 0.3 μl seeding additive in a second round of crystallization performed at 4 °C with 1 μl M1^{BC1}–FgD at 4 mg ml^{-1} and 1 μl of 16% PEG 3350 and 0.2 M sodium tartrate. Crystals were cryo-protected in the mother liquor solution supplemented with 25% ethylene glycol and flash cooled in liquid N_2. A native FgD–M1^{BC1} data set was recorded at 1.033 Å wavelength to 3.3 Å resolution limit at the Advanced Photon Source (APS, 23-ID-B, Argonne). Data were processed with the programs MOSFLM[26] and SCALA[33].

For crystallization of M1^{A}–FgD, M1^{A} was mixed with FgD and purified as described above. Crystals were grown using M1^{A}–FgD at 10.7 mg ml^{-1} and 1.3 M ammonium tartrate, 0.1 M MES, pH 6.25 as a precipitant. Crystals were cryo-protected in the mother liquor solution supplemented with 25% glycerol and flash cooled in liquid N_2. A native data set for crystals of M1^{A}–FgD was recorded at 0.9800 Å wavelength to 7.5 Å resolution at APS (23-ID-B), and data were processed as described above.

The diffraction data were truncated at resolution limits of 3.3 Å (I/σ_I of 1.4) and 7.5 Å (I/σ_I of 1.8) for M1^{BC1}–FgD and M1^{A}–FgD, respectively, as suggested by an analysis of σ_A values as a function of resolution[34].

Structure determination and refinement. Phases for M1^{BC1}–FgD and M1^{A}–FgD were determined by molecular replacement using the program Phaser[28] and the structure of human FgD[14] (PDB code 3E1I) as a search model. Four FgD molecules were identified through molecular replacement to occupy the asymmetric unit of M1^{BC1}–FgD (final rotation function and translation function *Z*-scores of 12.2 and 69.7, and initial R_{work}, R_{free} of 41.6%, 41.2%). Two FgD molecules were identified through molecular replacement to occupy the asymmetric unit of M1^{A}–FgD (final rotation function and translation function *Z*-scores of 5.2 and 25.5, and initial R_{work}, R_{free} of 44.1%, 42.1%). A single M1^{BC1}–FgD complex occupied the asymmetric unit of its crystal, and a single M1^{A}–FgD complex occupied the asymmetric unit of its crystal.

Continuous electron density corresponding to the backbone of M1^{BC1} in M1^{BC1}–FgD was evident in initial electron density maps, with sufficient side-chain density for a tentative register to be assigned. The register was verified by independent means as follows. Ile 148 in M1^{BC1} was substituted with methionine, and the resulting mutant protein was biosynthetically labelled with selenomethionine and crystallized in complex with FgD, as above. A highly redundant (1,080°) data set at the selenium anomalous absorption peak (0.9795 Å wavelength) was recorded at the APS (23-ID-D) to increase the signal to noise ratio. Data were processed using HKL2000[27]. A difference anomalous map was calculated using FFT[33], and two anomalous peaks were located, resulting in the unambiguous location of residue 148 on each chain of the M1^{BC1} coiled coil (Supplementary Fig. 2). The entirety of M1^{BC1} visible in the crystal structure was α-helical, and thus specification of the position of residue 148 enabled assignment of the remaining residues.

Refinement of the M1^{BC1}–FgD model was performed using CNS[35] and Refmac[33]. All B-factors were initially set to 90 Å^2 and subsequently refined as side-chain and main-chain groups using bgroup from CNS[36]. Four-fold non-crystallographic symmetry (NCS) restraints were applied to FgD (medium restraints for the main chain and loose restraints for the side chain); each of the three chains of the fibrinogen αβγ heterotrimer formed a separate NCS group. The model was then refined using Refmac5, with five alternating macro-cycles of model building and refinement. Model building was guided by inspection of σ_A-weighted $2F_o - F_c$ and $F_o - F_c$ omit maps[37], and was carried out using COOT[38]. Side chains were modelled as preferred rotamers. Each refinement macro-cycle consisted of 10 micro-cycles of maximum likelihood restrained refinement using standard parameters, except for the following adjustments. Owing to the moderate resolution of the data, a weighting term of 0.01 was used to favour geometric restraints, and an overall temperature factor model and a Babinet scaling model were used[34].

Structure validation was performed using Procheck[39] and Molprobity[40]. In the final M1^{BC1}–FgD model, 97.8% and 99.3% of residues were in allowed and generously allowed Ramachandran regions, respectively. The final map had correlation coefficients of 0.90 and 0.68 for the main chain and side chains, respectively, as calculated with OVERLAPMAP[33]. The Molprobity[40] clash score was 24.63 (89th percentile) and overall score was 3.18 (77th percentile).

Molecular figures were generated with PyMol (http://pymol.sourceforge.net).

Modelling of M1–fibrinogen. The A region of intact M1 was modelled based on the structure of M1^{AB} (ref. 8), and the B repeats, S region and the initial few residues of the C repeats were modelled based on the structure of M1^{BC1} from M1^{BC1}–FgD. For C-terminal portions of M1 for which no structural information exists, an α-helical dimeric coiled coil was modelled. Fibrinogen was modelled using the structure of chicken fibrinogen[41], which has been determined to a higher resolution limit than human fibrinogen[21]. The human and chicken fibrinogen structures show a similar organization, with some flexibility between the FgD portions due to a bend in the central fragment E portion. The initial M1–fibrinogen model, which has intact M1 at the centre of a cross-like structure formed by four fibrinogen molecules, was enlarged as follows. Because each of the four fibrinogen molecules has a second binding site for M1, M1 was modelled at these second sites based on the structure of M1^{BC1}–FgD. Furthermore, because M1 binds four fibrinogen molecules, three more fibrinogen molecules were modelled at these second sites based on the structure of M1^{BC1}–FgD. This procedure was carried on iteratively to yield the model of the M1–fibrinogen network. Similar procedures were carried out for M1 mutants.

Co-precipitation assays. Ten micrograms of wild-type or variant M1 proteins were mixed with 20 μg of FgD in 50 μl of binding buffer (300 mM NaCl, 50 mM sodium phosphate buffer, pH 8.0, 50 mM imidazole, 0.1% (v/v) Triton X-100) at 37 °C for 30 min. Twenty microlitres of Ni^{2+}-NTA agarose beads were equilibrated in binding buffer and then added to the protein mix and incubated for 30 min at 37 °C under agitation. The beads were washed three times in 200 μl of binding buffer and eluted by boiling the beads for 5 min in non-reducing 5× SDS–PAGE sample loading buffer. Fractions corresponding to unbound and bound proteins were resolved by non-reducing SDS–PAGE.

A similar procedure was used for human IgG Fc fragment (Calbiochem), except for the following changes: 25 μg of M1 or M1*-R protein and 100 μg of Fc were mixed; this mixture was added to 50 μl of equilibrated Ni^{2+}-NTA agarose beads; washes were 1 ml each; and reducing SDS–PAGE sample loading buffer was used.

Heparin binding protein immunoblot. Human neutrophils were purified from healthy donors blood using the PolymorphPrep system (Axis-Shield). Thirteen million neutrophils were incubated with 86 μg ml^{-1} wild-type or variant M1 protein at 37 °C for 30 min. After centrifugation, the supernatants were resolved on a 12% SDS–PAGE reducing gel and transferred to a PVDF membrane for immunoblotting. Recombinant human heparin binding protein (HBP) (R&D Systems) was used as a reference. The membrane was blocked with Tris buffered saline containing 5% non-fat milk or 5% bovine serum albumin (BSA). HBP was detected using either a primary rabbit anti-human-HBP polyclonal antibody (Sigma) or a mouse anti-human-HBP monoclonal antibody (R&D), peroxidase-conjugated secondary antibodies (Santa Cruz), and the ECL system (Pierce).

Electron microscopy. Mixtures containing 10 μg of M proteins and 100 μg of human fibrinogen (Calbiochem) were incubated at 37 °C for 15 min before being

©2011 Macmillan Publishers Limited. All rights reserved

deposited on glow-discharged copper grids coated with carbon film (Electron microscopy sciences, catalogue number CF300-Cu) for 2 min. The grids were subsequently washed twice with water and negatively stained with 0.2% uranyl acetate for 45 s. Images were recorded using a FEI Tecnai F20 Twin transmission electron microscope at an accelerating voltage of 120 kV.

29. Kansal, R. G., McGeer, A., Low, D. E., Norrby-Teglund, A. & Kotb, M. Inverse relation between disease severity and expression of the streptococcal cysteine protease, SpeB, among clonal M1T1 isolates recovered from invasive group A streptococcal infection cases. *Infect. Immun.* **68,** 6362–6369 (2000).
30. Geiser, M., Cebe, R., Drewello, D. & Schmitz, R. Integration of PCR fragments at any specific site within cloning vectors without the use of restriction enzymes and DNA ligase. *Biotechniques* **31,** 88–90 92 (2001).
31. Doublié, S. Preparation of selenomethionyl proteins for phase determination in *Methods in Enzymology*, **276,** 523–530 (Academic, 1997).
32. Everse, S. J., Pelletier, H. & Doolittle, R. F. Crystallization of fragment D from human fibrinogen. *Protein Sci.* **4,** 1013–1016 (1995).
33. CCP4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
34. DeLaBarre, B. & Brunger, A. T. Considerations for the refinement of low-resolution crystal structures. *Acta Crystallogr. D* **62,** 923–932 (2006).
35. Brünger, A. Crystallography & NMR system: a new software for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
36. Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355,** 472–475 (1992).
37. Brünger, A. T., Adams, P. D. & Rice, L. M. New applications of simulated annealing in X-ray crystallography and solution NMR. *Structure* **5,** 325–336 (1997).
38. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
39. Laskowski, R. A., Moss, D. S. & Thornton, J. M. Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.* **231,** 1049–1067 (1993).
40. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
41. Yang, Z., Kollman, J. M., Pandi, L. & Doolittle, R. F. Crystal structure of native chicken fibrinogen at 2.7 Å resolution. *Biochemistry* **40,** 12515–12523 (2001).

©2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09933

Electromagnetically induced transparency and slow light with optomechanics

A. H. Safavi-Naeini[1]*, T. P. Mayer Alegre[1]*, J. Chan[1], M. Eichenfield[1], M. Winger[1], Q. Lin[1], J. T. Hill[1], D. E. Chang[2,3] & O. Painter[1]

Controlling the interaction between localized optical and mechanical excitations has recently become possible following advances in micro- and nanofabrication techniques[1,2]. So far, most experimental studies of optomechanics have focused on measurement and control of the mechanical subsystem through its interaction with optics, and have led to the experimental demonstration of dynamical back-action cooling and optical rigidity of the mechanical system[1,3]. Conversely, the optical response of these systems is also modified in the presence of mechanical interactions, leading to effects such as electromagnetically induced transparency[4] (EIT) and parametric normal-mode splitting[5]. In atomic systems, studies[6,7] of slow and stopped light (applicable to modern optical networks[8] and future quantum networks[9]) have thrust EIT to the forefront of experimental study during the past two decades. Here we demonstrate EIT and tunable optical delays in a nanoscale optomechanical crystal, using the optomechanical nonlinearity to control the velocity of light by way of engineered photon–phonon interactions. Our device is fabricated by simply etching holes into a thin film of silicon. At low temperature (8.7 kelvin), we report an optically tunable delay of 50 nanoseconds with near-unity optical transparency, and superluminal light with a 1.4 microsecond signal advance. These results, while indicating significant progress towards an integrated quantum optomechanical memory[10], are also relevant to classical signal processing applications. Measurements at room temperature in the analogous regime of electromagnetically induced absorption show the utility of these chip-scale optomechanical systems for optical buffering, amplification, and filtering of microwave-over-optical signals.

It is by now well known that the optical properties of matter can be dramatically modified by using a secondary light beam, approximately resonant with an internal process of the material system. As an example, an opaque object can be made transparent in the presence of a control beam; this is the phenomenon of EIT. A remarkable feature of EIT is the drastic reduction in the group velocity of light passing through the material, achieved inside a practically lossless transparency window. This aspect of the effect has been used in schemes whereby light may be slowed and stopped, making it an important building block in quantum information and communication proposals, as well as of great practical interest in classical optics and photonics. A simple upper-bound for the storage time in EIT-based proposals is the lifetime related to the internal processes of the material. These lifetimes can be extremely long in atomic gases, with storage times of the order of one second having been demonstrated[11] in Bose–Einstein condensates. Part of the vision for future scalable quantum networks has involved extending the remarkable results achieved in atomic experiments to a more readily deployable domain.

In the solid state, EIT has been demonstrated in quantum wells, dots and nitrogen–vacancy centres[12–14]. But the fast dephasing rates and inhomogeneous broadening of solid-state electronic resonances have led to a plethora of other methods and techniques. Elegant experiments with stimulated Brillouin scattering in fibres[15] and coherent population oscillations[16] have been used to delay intense classical light. Alternatively, for quantum storage and buffering, techniques related to photon-echo spectroscopy (for example, controllable reversible inhomogeneous broadening[17] and atomic frequency combs[18]) have been used successfully to achieve solid-state quantum memories. In chip-scale photonics, dynamically tunable arrays of cavities, displaying EIT, are an intriguing analogue to ensembles of atoms and provide a route to slowing and stopping light all-optically[19]. Generally, the elements in the arrays have consisted of coupled optical or plasmonic resonances, and have been demonstrated with couplings engineered to give rise to Fano-like interference[20]. However, a significant limitation in these all-photonic systems is the short optical resonance lifetime. We demonstrate here that in addition to optically controlled switching of a probe beam, as recently presented by others[4], EIT in an optomechanical cavity may be used to change the group velocity of light significantly. As such, tunable optical delay, with delay times limited by the much longer mechanical resonance lifetime of the optomechanical system, may be achieved. These delays are also attainable across a broad spectrum of wavelengths; indeed, recent circuit cavity electromechanics experiments in the strong-coupling regime have demonstrated EIT and group velocity control at microwave frequencies[21]. Additionally, the ability to create arrays of coupled devices by 'printing' optomechanical circuits onto a Si microchip[22–24], for example, allows one to create a much larger delay-bandwidth product (scaling as $\sqrt{N}$, N being the number of cavity elements)[10]; such arrays provide a means to sample an incoming optical pulse shape, store it and retrieve it, much like an ensemble of atoms does in atomic EIT.

EIT in optomechanical systems can be understood physically as follows. The conventional radiation pressure interaction between a near-resonant cavity light field and mechanical motion is modelled by the nonlinear Hamiltonian $H_{\mathrm{int}} = \hbar g \hat{a}^\dagger \hat{a} \left(\hat{b} + \hat{b}^\dagger \right)$. Here $\hat{a}$ ($\hat{a}^\dagger$) and $\hat{b}$ ($\hat{b}^\dagger$) are the annihilation (creation) operators of photon and phonon resonator quanta, respectively, g is the optomechanical coupling rate corresponding physically to the shift in the optical mode's frequency due to the zero-point fluctuations of the phonon mode, and $\hbar$ is $h/2\pi$, where h is Planck's constant. By driving the system with an intense red-detuned optical 'control' beam at frequency ω_c, as shown in Fig. 1a, the form of the effective interaction changes (in the resolved sideband limit) to that of a beam-splitter-like Hamiltonian, $H_{\mathrm{int}} = \hbar G \left(\hat{a}^\dagger \hat{b} + \hat{a} \hat{b}^\dagger \right)$. Here, the zero-point-motion coupling rate g is replaced by a much stronger parametric coupling rate $G = g\sqrt{\langle n_c \rangle}$ between light and mechanics, where $\langle n_c \rangle$ is the stored intracavity photon number induced by the control beam. Viewed in a dressed-state picture, with the control beam detuning from the optical cavity resonance (ω_o) set equal to the mechanical frequency (ω_m), $\Delta_{\mathrm{OC}} \equiv \omega_o - \omega_c \approx \omega_m$, the optical and mechanical modes $\hat{a}$ and $\hat{b}$ become coupled (denoted $\hat{a}_d$ and $\hat{b}_d$ in Fig. 1b). The dressed mechanical mode, now effectively a phonon–photon polariton, takes on a weakly photonic nature, coupling it to the optical loss channels at a rate $\gamma_{\mathrm{om}} \equiv C\gamma_i$, where the optomechanical

[1] Thomas J. Watson Sr Laboratory of Applied Physics, California Institute of Technology, Pasadena, California 91125, USA. [2] Institute for Quantum Information, California Institute of Technology, Pasadena, California 91125, USA. [3] Center for the Physics of Information, California Institute of Technology, Pasadena, California 91125, USA.
*These authors contributed equally to this work.

©2011 Macmillan Publishers Limited. All rights reserved

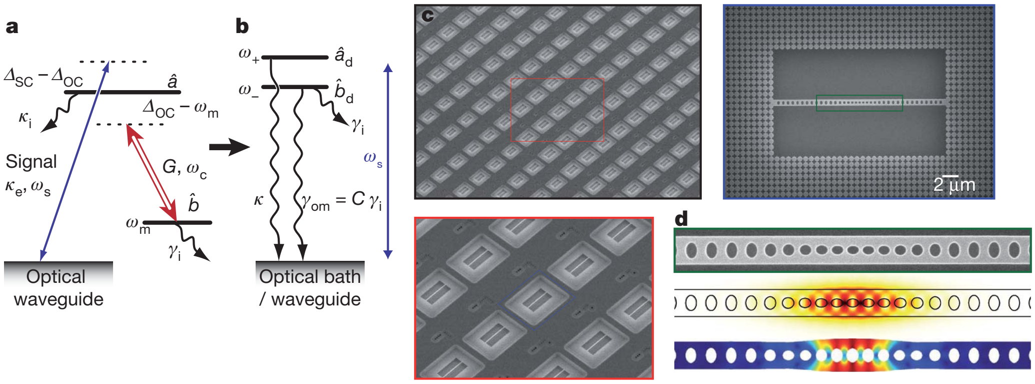


Figure 1 | Optomechanical system. a, Level-diagram picture, showing three 'levels' that represent the optical mode $\hat{a}$, the mechanical mode $\hat{b}$ and the 'bath' of optical waveguide modes. The transitions between modes driven by the signal and control beams are indicated by blue and red double-headed arrows, respectively. Wavy black arrows indicate decay from the different modes. See text for definitions of symbols in **a** and **b**. **b**, The control beam at ω_c drives the transition between the optical and mechanical mode, dressing the optical and mechanical modes, resulting in the dressed state picture with dressed modes $\hat{a}_d$ and $\hat{b}_d$. **c**, Series of scanning electron micrographs, showing large array of optomechanical crystal nanocavities (top-left panel), zoomed-in image of device array (bottom-left panel), and zoomed-in image of top-view of single cavity device (top-right panel). **d**, From top to bottom: scanning electron micrograph of a zoomed-in region showing the OMC defect region; finite-element-method (FEM) simulation results for the optical field showing the electrical field intensity $|\mathbf{E}(\mathbf{r})|$; FEM-simulated mechanical mode with the total displacement $|\mathbf{Q}(\mathbf{r})|$ shown.

cooperativity is defined as $C \equiv 4G^2/\kappa\gamma_i$ for an optical cavity decay rate of κ, and an intrinsic mechanical resonance damping rate of γ_i.

The drive-dependent loss rate γ_{om} has been viewed in most previous studies as an incoherent, quantum-limited loss channel, and was used in recent experiments to cool the mechanical resonator close to its quantum ground state[25]. In the dressed mode picture, by analogy to the dressed state view of EIT[7], it becomes clear that a coherent cancellation of the loss channels in the dressed optical and mechanical modes is possible, and can be used to switch the system from absorptive to transmittive in a narrow band around cavity resonance. Much as in atomic EIT, this effect causes an extremely steep dispersion for the transmitted probe photons, with a group delay on resonance of (see Supplementary Information)

$$\tau^{(T)}|_{\omega=\omega_m} = \frac{2}{\gamma_i}\frac{(\kappa_e/\kappa)C}{(1+C)(1-(\kappa_e/\kappa)+C)} \qquad (1)$$

where κ_e is the optical coupling rate between the external optical waveguide and the optical cavity, and the delay is dynamically tunable via the control beam intensity through C.

Nano- and micro-optomechanical resonators take a variety of forms, among which optomechanical crystals (OMC) have been used to demonstrate large radiation-pressure-induced interaction strengths between gigahertz mechanical resonances and near-infrared optical resonances[24]. The nanobeam OMC cavity used in this study (Fig. 1c and d) uses a periodic free-standing Si structure to create high-Q co-localized optical and mechanical resonances. These devices can be printed and etched into the surface of a Si chip in large arrays (Fig. 1c), and are designed to operate optically in the telecommunications band ($\lambda_o = 1{,}550$ nm) and acoustically at microwave frequencies ($\omega_m/2\pi = 3.75$ GHz). The theoretical optomechanical coupling rate g between co-localized photon and phonon modes is $g/2\pi \approx 800$ kHz. By optimizing the arrangement of holes in the central cavity region of the nanobeam where light and sound are localized, an intrinsic optical decay rate of $\kappa_i/2\pi \approx 290$ MHz is obtained for the optical cavity mode, placing the optomechanical system in the resolved sideband regime ($\omega_m/\kappa_i \gg 1$) necessary for EIT. The corresponding mechanical resonance is measured to have an intrinsic damping rate of $\gamma_i/2\pi \approx 250$ kHz at temperature $T = 8.7$ K, corresponding to a mechanical Q-factor of $Q_m = 15{,}000$. Light is coupled into and out of the device using a specially prepared optical fibre taper, which when placed in the near-field of the nanobeam cavity couples the guided modes of the taper evanescently to the optical resonances of the nanobeam (see Supplementary Information for details of the optical cavity loading).

In order to characterize the near-resonance optical reflection of the cavity system, a sideband of the control beam is created using electro-optic modulation (see Methods and Supplementary Information), forming a weak signal beam with tunable frequency ω_s. The results of measurements performed at a cryogenic temperature of 8.7 K are shown in Fig. 2. Here, the control beam laser power was varied from 6 µW ($\langle n_c \rangle = 25$) to nearly 250 µW ($\langle n_c \rangle = 1{,}040$). The frequencies of both the control and signal beams are swept in order to map out the system dependence on control beam detuning, Δ_{OC}, and the two-photon detuning, $\Delta_{SC} = \omega_s - \omega_c$. The resulting reflected optical signal intensity, separated from the control beam via a modulation and lock-in technique (see Methods and Supplementary Information), is shown in Fig. 2a for a series of control laser detunings. Visible in each of the plots is a broad resonance corresponding to the bare optical cavity response with loaded linewidth $\kappa/2\pi \approx 900$ MHz. A much narrower reflection dip feature, corresponding to the transparency window, can also be seen near the cavity line centre. The position of the narrow reflection dip tracks with a two-photon detuning equal to the mechanical resonance frequency, $\Delta_{SC} \approx \omega_m$. This region is shown in more detail in Fig. 2b, where the Fano-like structure of the optical response is apparent. Each curve in Fig. 2a and b is a horizontal slice of the data presented in Fig. 2c, where the reflectivity is plotted as a function of both ω_c and Δ_{SC}. The transparency window is shown to be fully controllable via the applied light field, the window expanding and contracting with the control beam laser power (Fig. 2d and e). At the maximum stable control power (unstable regions due to a thermo-optic bistability induced by optical absorption are shown as hatched regions in Fig. 2c), a transparency window approaching 5 MHz is obtained.

Model fits to the reflection spectra (see Supplementary Information) are shown as solid curves in Fig. 2a and b. The resulting fitted values for $\gamma_{om} = 4G^2/\kappa$ for each control power are shown in Fig. 2e. A linear fit to the extracted data yields a value for the zero-point-motion coupling constant of $g/2\pi = 800$ kHz, in agreement with the value obtained from independent optical transduction measurements of the thermal Brownian motion of the mechanical oscillator[24]. In addition to the intensity response of the optomechanical cavity there is the phase response, which provides a measure of the group delay of the modulated optical signal beam as it passes through the cavity. For the 89-kHz modulation of the signal beam used in our experiments, corresponding to a free-space signal wavelength of ~3.4 km, phase shifts between the modulation sidebands and the signal carrier on the order of a fraction of a radian are measured in the region where Δ_{SC} is within a

©2011 Macmillan Publishers Limited. All rights reserved

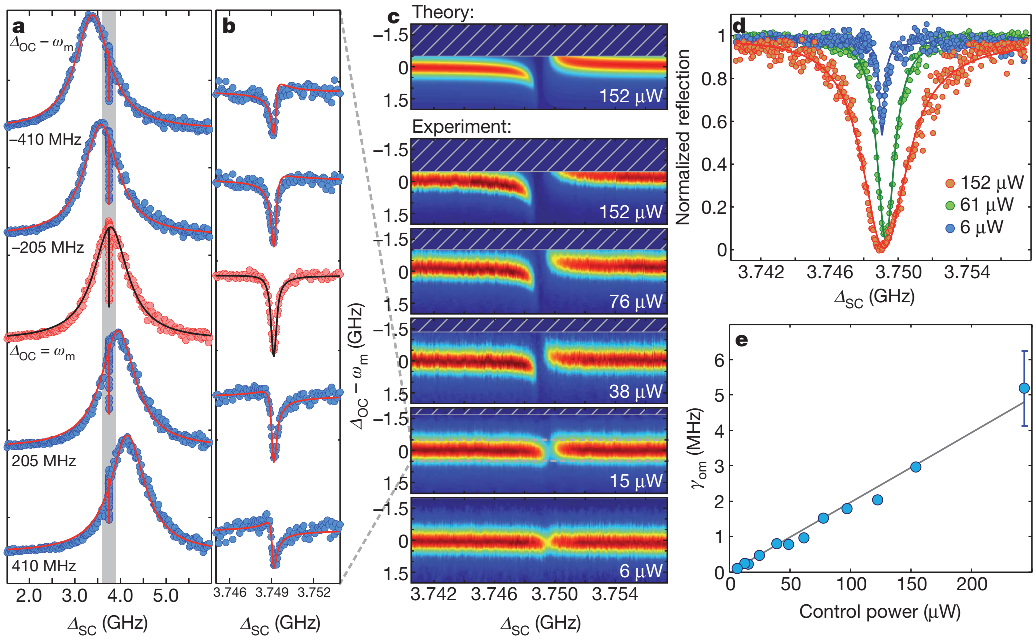


Figure 2 | Optical reflection response at temperature $T = 8.7$ K. a, Measured normalized reflection (dots) of the signal beam as a function of the two-photon detuning (Δ_{SC}) for a control beam power of 15 μW. **b**, Zoom-in of the reflected signal about the transparency window. Each spectrum in **a** and **b** corresponds to a different control laser detuning ($\Delta_{OC}-\omega_m$) as indicated. Solid curves correspond to model fits to the data (see Supplementary Information). **c**, Intensity plots for the signal beam reflection as a function of both control laser detuning (Δ_{OC}) and two-photon detuning (Δ_{SC}) for a series of different control beam powers (as indicated). The hatched areas are unstable regions for the control laser detuning at the given input power. The top plot is the theoretically predicted reflection spectrum for the highest control beam power. **d**, Transparency window versus control beam power for control laser detuning $\Delta_{OC}\approx\omega_m$. **e**, Transparency window bandwidth ($\gamma_{om}=4G^2/\kappa$) versus control beam power (error bars indicate the standard deviation in the fit of γ_{om} to the EIT intensity spectra versus Δ_{OC}). The solid line represents the bandwidth scaling for a single best-fit value of $g/2\pi = 800$ kHz.

mechanical linewidth of ω_m. The measured phase-shifts for the reflected signal beam correspond to advances in time of the modulated signal, pointing to causality-preserving superluminal effects. A plot of the peak effective signal advance versus control beam power is plotted in Fig. 3a, ascertained from a fit to the reflection phase response spectra (Fig. 3b). For the highest control power, the reflected signal is advanced by 1.3 μs, roughly 7,000 times longer than the bare optical cavity lifetime.

The delay in transmission is directly related to the advance on reflection through the bare cavity transmission contrast (measured independently; see Supplementary Information). As such, we plot the corresponding transmission group delay of the signal in Fig. 3c. The theoretical delay/advance of the modulated signal beam for system parameters given by fits to the EIT intensity spectra are shown as dashed curves in Fig. 3a and c, indicating good agreement with the measured phase response. As can be seen in these data, the maximum measured transmission delay is $\tau^{(T)}\approx 50$ ns, which—although corresponding to significant slowing of light (to a velocity of $v_g\approx 40\,\mathrm{m\,s^{-1}}$) through the few-micrometre-long structure—is much smaller than the measured reflected signal advance or the limit set by the intrinsic mechanical damping ($2/\gamma_i\approx 1.4$ μs). This is due to the weak loading of the optical cavity in these experiments (see Supplementary Information), and the resulting small fraction of transmitted light that actually passes through the cavity.

In addition to the observed EIT-like behaviour of the optomechanical system, a similar phenomena to that of electromagnetically induced absorption (EIA)[26] in atomic systems can be realized by setting the detuning of the control beam to the blue side of the optomechanical cavity resonance ($\Delta_{OC}<0$). Under blue-detuned pumping, the effective Hamiltonian for the optical signal and mechanical phonon mode becomes one of parametric amplification, $H_{\mathrm{int}}=\hbar G\left(\hat{a}^{\dagger}\hat{b}^{\dagger}+\hat{a}\hat{b}\right)$. The measured reflection spectrum from the OMC is shown in Fig. 3d, where the reflectivity of the cavity system is seen to be enhanced around the two-photon detuning $\Delta_{SC}\approx\omega_m$, a result of the increased 'absorption' (feeding) of photons into the cavity. As discussed further in the Supplementary Information, at even higher control beam powers such that $C\gtrsim 1$, the system switches from EIA to parametric amplification, resulting in optical signal amplification, and eventually phonon-lasing.

Reflection spectroscopy at room temperature (296 K) of the optomechanical cavity has also been performed (presented in Supplementary Information), and yields similar results to that of the cryogenic measurements, albeit with a larger value of $\langle n_c\rangle$ required to reach a given cooperativity (see Fig. 3e) owing to the larger intrinsic mechanical dissipation at room temperature ($\gamma_i = 2\pi\times 1.9$ MHz). Beyond the initial demonstrations of EIT and EIA behaviour in the OMC cavities presented here, it is fruitful to consider the bandwidth and signal delay limits that might be attainable with future improvements in device material or geometry. For instance, the transparency bandwidth of the current devices is limited by two-photon absorption of the control beam in the silicon cavities; a move to larger-bandgap dielectric materials, such as silicon nitride, should allow intra-cavity photon numbers of 10^6 (limited by linear material absorption), resulting in a transparency window approaching $G=g\sqrt{\langle n_c\rangle}\approx 2\pi(1\text{ GHz})$. Also, recent research into low-loss GHz mechanic resonators[27] should enable slow light optical delays approaching 10 μs at room temperature, roughly equivalent to the optical path length of a kilometre of optical fibre. Much like the acoustic wave devices used in electronic systems[28], optomechanical devices with these attributes would enable chip-scale microwave photonic systems capable of advanced signal processing in the optical domain, such as that needed for emerging broadband wireless access networks or more specialized applications, such as true-time delays in radar systems[8].

The limiting factor for quantum applications of optomechanical systems is the re-thermalization time of the mechanical resonator, $\tau_{th}=\hbar Q_m/kT$, which in the case of a quantum optical memory

©2011 Macmillan Publishers Limited. All rights reserved

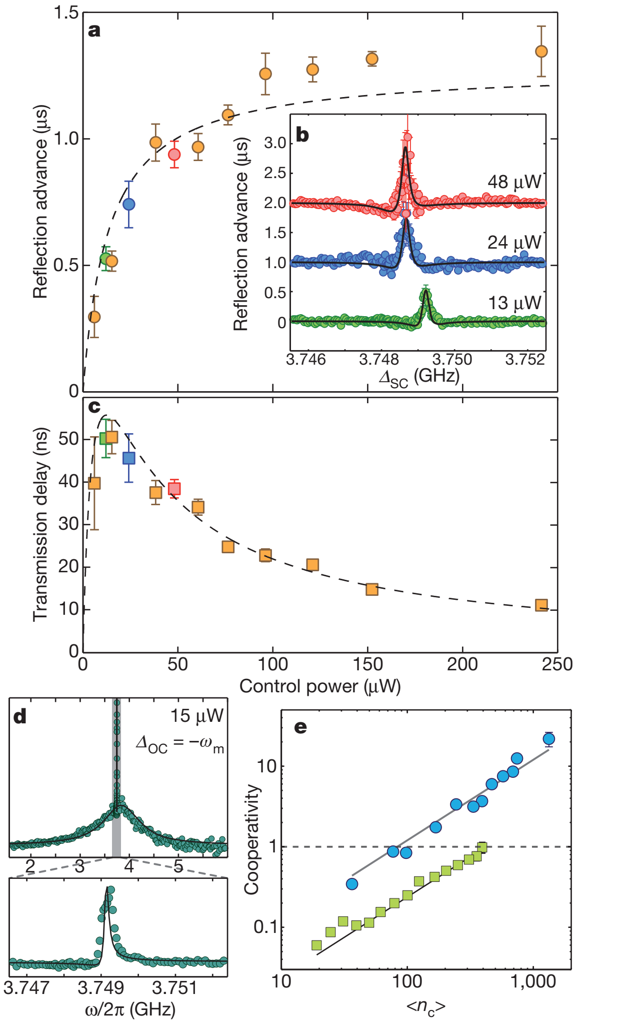


Figure 3 | Measured temporal shifts and amplification. **a**, Maximum measured reflected signal advance as a function of the control beam power. **b**, Measured reflected signal advance versus two-photon detuning, Δ_{SC}. Solid curves correspond to fit from model (see Supplementary Information). Curves at different control powers are shifted for clarity. **c**, Inferred maximum transmitted signal delay versus control beam power. Dashed lines in **a** and **c** are theoretical advance/delay times predicted from model of optomechanical system based on intensity response of the optomechanical system. **d**, Measured signal reflection as a function of two-photon detuning for the control laser blue-detuned from the cavity. **e**, Measured cooperativity for sample temperature of 296 K (squares) and 8.7 K (circles) as a function of the average number of control photons inside the cavity. Error bars indicate the standard deviation in the model fit to the EIT spectral data at each control beam power.

represents the average storage time of a single photon before excitation of the system by a thermal bath phonon. For the devices studied here, despite the optical cooling and reduced phonon occupancy of the mechanical resonator provided by the control beam (the cooling rate being equal to the transparency window bandwidth[10]), the re-thermalization time is limited to $\tau_{th} \approx 12$ ns by the 8.7 K bath temperature. Reducing the operating temperature further to a value below 100 mK (routinely attained in a dilution refrigerator) would not only increase the re-thermalization time through a lower bath temperature, but should also result in a significant increase in the mechanical Q-factor. Taken together, the resulting re-thermalization time in the current OMC devices at $T = 100$ mK is likely to be of the order of 100 μs, which although not nearly as long as what has been achieved in atomic systems[11], still represents a substantial storage time compared to the realizable GHz bandwidth of the system. Additionally, optomechanical processes similar to the EIT behaviour measured here have also been proposed[29,30] to provide an optical interface between, for instance, atomic and superconducting circuit quantum systems, enabling the formation of hybrid quantum networks.

METHODS SUMMARY

Fabrication. The nanobeam cavities were fabricated using a silicon-on-insulator wafer from SOITEC (resistivity $\rho = 4$–$20\,\Omega$ cm, device layer thickness $t = 220$ nm, buried-oxide layer thickness 2 μm). The cavity geometry is defined by electron beam lithography followed by inductively-coupled-plasma reactive ion etching to transfer the pattern through the 220-nm silicon device layer. The cavities were then undercut using an HF:H_2O solution to remove the buried oxide layer, and cleaned using a piranha etch/HF etch cycle. The dimensions and design of the nanobeam will be discussed in detail elsewhere.

Experimental set-up. We demonstrate EIT via reflection measurements of the optically pumped system at varying $\langle n_c \rangle$. Using the experimental set-up shown in Supplementary Information, a laser beam at ω_c (the control beam) is sent through an electro-optical modulator with drive frequency equal to Δ_{SC}, creating an optical sideband at frequency ω_s (the signal beam), which is amplitude modulated at $\omega_{LI}/2\pi = 89$ kHz. Since the control beam is detuned from the cavity by $|\Delta_{OC}| \gg \kappa$, it is effectively filtered when looking in reflection, while the modulated signal beam at $\omega_c \pm \Delta_{SC}$ (where the sign is that of Δ_{OC}) is near resonance with the optical cavity and is reflected. The reflected signal beam is detected using a 12-GHz New Focus PIN photo-diode, with the output electrical signal sent to a lock-in amplifier where the component related to the modulated tone ($\omega_{LI} = 89$ kHz) is amplified and sent to an oscilloscope. By scanning both the laser frequency ω_c and the two-photon detuning Δ_{SC}, a full map of the reflectivity is obtained. Additionally, by using a lock-in amplifier, the phase of the modulated signal sidebands relative to the carrier can be measured, giving a direct measurement of the group delay imparted on the optical signal beam by the optomechanical cavity.

Received 8 December 2010; accepted 21 February 2011.
Published online 16 March 2011.

1. Kippenberg, T. J. & Vahala, K. J. Cavity optomechanics: back-action at the mesoscale. *Science* **321,** 1172–1176 (2008).
2. Favero, I. & Karrai, K. Optomechanics of deformable optical cavities. *Nature Photon.* **3,** 201–205 (2009).
3. Braginsky, V. B. *Measurement of Weak Forces in Physics Experiments* (Univ. Chicago Press, 1977).
4. Weis, S. *et al.* Optomechanically induced transparency. *Science* **330,** 1520–1523 (2010).
5. Gröblacher, S., Hammerer, K., Vanner, M. & Aspelmeyer, M. Observation of strong coupling between a micromechanical resonator and an optical cavity field. *Nature* **460,** 724–727 (2009).
6. Hau, L. V., Harris, S. E., Dutton, Z. & Behroozi, C. H. Light speed reduction to 17 metres per second in an ultracold atomic gas. *Nature* **397,** 594–598 (1999).
7. Fleischhauer, M., Imamoglu, A. & Marangos, J. P. Electromagnetically induced transparency: optics in coherent media. *Rev. Mod. Phys.* **77,** 633–673 (2005).
8. Boyd, R. W. & Gauthier, D. J. Controlling the velocity of light pulses. *Science* **326,** 1074–1077 (2009).
9. Kimble, H. J. The quantum internet. *Nature* **453,** 1023–1030 (2008).
10. Chang, D., Safavi-Naeini, A. H., Hafezi, M. & Painter, O. Slowing and stopping light using an optomechanical crystal array. *N. J. Phys.* **13,** 023003 (2011).
11. Zhang, R., Garner, S. R. & Hau, L. V. Creation of long-term coherent optical memory via controlled nonlinear interactions in Bose-Einstein condensates. *Phys. Rev. Lett.* **103,** 233602 (2009).
12. Phillips, M. C. *et al.* Electromagnetically induced transparency in semiconductors via biexciton coherence. *Phys. Rev. Lett.* **91,** 183602 (2003).
13. Santori, C. *et al.* Coherent population trapping of single spins in diamond under optical excitation. *Phys. Rev. Lett.* **97,** 247401 (2006).
14. Xu, X. *et al.* Coherent population trapping of an electron spin in a single negatively charged quantum dot. *Nature Phys.* **4,** 692–695 (2008).
15. Thévenaz, L. Slow and fast light in optical fibres. *Nature Photon.* **2,** 474–481 (2008).
16. Bigelow, M. S., Lepeshkin, N. N. & Boyd, R. W. Superluminal and slow light propagation in a room-temperature solid. *Science* **301,** 200–202 (2003).
17. Afzelius, M., Simon, C., de Riedmatten, H. & Gisin, N. Multimode quantum memory based on atomic frequency combs. *Phys. Rev. A* **79,** 052329 (2009).
18. de Riedmatten, H., Afzelius, M., Staudt, M. U., Simon, C. & Gisin, N. A solid-state light-matter interface at the single-photon level. *Nature* **456,** 773–777 (2008).
19. Yanik, M. F., Suh, W., Wang, Z. & Fan, S. Stopping light in a waveguide with an all-optical analog of electromagnetically induced transparency. *Phys. Rev. Lett.* **93,** 233903 (2004).
20. Xu, Q. *et al.* Experimental realization of an on-chip all-optical analogue to electromagnetically induced transparency. *Phys. Rev. Lett.* **96,** 123901 (2006).
21. Teufel, J. D. *et al.* Circuit cavity electromechanics in the strong coupling regime. *Nature* doi:10.1038/nature09898 (10 March 2011); preprint at 〈http://arXiv.org/abs/1011.3067〉 (2010).
22. Notomi, M., Kuramochi, E. & Tanabe, T. Large-scale arrays of ultrahigh-Q coupled nanocavities. *Nature Photon.* **2,** 741–747 (2008).

©2011 Macmillan Publishers Limited. All rights reserved

23. Li, M. *et al.* Harnessing optical forces in integrated photonic circuits. *Nature* **456,** 480–484 (2008).
24. Eichenfield, M., Chan, J., Camacho, R. M., Vahala, K. J. & Painter, O. Optomechanical crystals. *Nature* **462,** 78–82 (2009).
25. Rocheleau, T. *et al.* Preparation and detection of a mechanical resonator near the ground state of motion. *Nature* **463,** 72–75 (2010); published online 9 December 2009.
26. Lezama, A., Barreiro, S. & Akulshin, A. M. Electromagnetically induced absorption. *Phys. Rev. A* **59,** 4732–4735 (1999).
27. Nguyen, C. T.-C. MEMS technology for timing and frequency control. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **54,** 251–270 (2007).
28. Lakin, K., Kline, G. & McCarron, K. Development of miniature filters for wireless applications. *IEEE Trans. Microwave Theory Techn.* **43,** 2933–2939 (1995).
29. Stannigel, K., Rabl, P., Srensen, A. S., Zoller, P. & Lukin, M. D. Optomechanical transducers for long-distance quantum communication. *Phys. Rev. Lett.* **105,** 220501 (2010).
30. Safavi-Naeini, A. H. & Painter, O. Proposal for an optomechanical traveling wave phonon-photon translator. *N. J. Phys.* **13,** 013017 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Schwab for providing the microwave modulation source used in this work. This work was supported by the DARPA/MTO ORCHID programme through a grant from AFOSR, and the Kavli Nanoscience Institute at Caltech. A.H.S.-N. and J.C. acknowledge support from NSERC.

Author Contributions J.C., A.H.S.-N. and M.E. performed the device design, and J.C. performed the device fabrication with support from M.W. and J.T.H. Measurements and data analysis were performed by A.H.S.-N. and T.P.M.A., with support from both D.E.C. and Q.L. and supervision by O.P. All authors contributed to the writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to O.P. (opainter@caltech.edu).

©2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09979

High-frequency, scaled graphene transistors on diamond-like carbon

Yanqing Wu[1], Yu-ming Lin[1], Ageeth A. Bol[1], Keith A. Jenkins[1], Fengnian Xia[1], Damon B. Farmer[1], Yu Zhu[1] & Phaedon Avouris[1]

Owing to its high carrier mobility and saturation velocity, graphene has attracted enormous attention in recent years[1–5]. In particular, high-performance graphene transistors for radio-frequency (r.f.) applications are of great interest[6–13]. Synthesis of large-scale graphene sheets of high quality and at low cost has been demonstrated using chemical vapour deposition (CVD) methods[14]. However, very few studies have been performed on the scaling behaviour of transistors made from CVD graphene for r.f. applications, which hold great potential for commercialization. Here we report the systematic study of top-gated CVD-graphene r.f. transistors with gate lengths scaled down to 40 nm, the shortest gate length demonstrated on graphene r.f. devices. The CVD graphene was grown on copper film and transferred to a wafer of diamond-like carbon. Cut-off frequencies as high as 155 GHz have been obtained for the 40-nm transistors, and the cut-off frequency was found to scale as 1/(gate length). Furthermore, we studied graphene r.f. transistors at cryogenic temperatures. Unlike conventional semiconductor devices where low-temperature performance is hampered by carrier freeze-out effects, the r.f. performance of our graphene devices exhibits little temperature dependence down to 4.3 K, providing a much larger operation window than is available for conventional devices.

Graphene has zero bandgap, and therefore devices fabricated from it have a small on-off ratio. Although bandgap engineering techniques—such as nano-ribbon fabrication or the application of a strong displacement field to bilayer graphene—have been developed to open a small bandgap in graphene[15–19], the development of a reliable technique to create a sizable gap without degrading the electronic properties of the material remains challenging. On the other hand, a large on-off ratio is not necessary for many r.f. applications, such as amplifiers or mixers, and significant progress has been made in the development of high-performance r.f. transistors based on graphene materials produced by different synthesis techniques. For example, a maximum cut-off frequency of 300 GHz has been obtained for devices based on exfoliated graphene (ref. 11), and a maximum cut-off frequency of 100 GHz for devices based on epitaxial graphene grown on silicon carbide (ref. 9). In modern electronics, large volume production and low cost are crucial properties for any new technology. It has been shown that growing graphene on a Cu substrate by CVD can produce large-size, high-quality sheets at low cost[14]. Previous studies on r.f. transistors made from transferred graphene typically used silicon dioxide (SiO_2) as the substrate. However, graphene devices fabricated on SiO_2 have been found to suffer from additional scattering associated with the low surface phonon energy (59 meV) and large trap density in SiO_2, resulting in deterioration of both device properties and uniformity across the wafer. To mitigate these problems, here we introduce a new substrate for graphene r.f. transistors, namely, a diamond-like carbon (DLC) film grown on SiO_2. Compared to SiO_2 and most other substrates, the DLC film has a higher phonon energy (owing to the high phonon energy in diamond (165 meV)) and a lower surface trap density (DLC is non-polar and chemically inert)[20]. These desirable properties help the high performance of graphene r.f. transistors to be achieved.

Single-layer graphene was grown on copper foil at high temperatures close to 1,000 °C. Using a polymethylmethacrylate (PMMA) film as a protecting layer, the graphene sheet formed on Cu was then freed by dissolving the Cu using a solution of $FeCl_3$. The transfer process was completed by transferring the PMMA-graphene to the DLC substrate and subsequently removing the PMMA. Raman spectroscopy was used to verify the single-layer nature of the graphene after the transfer (see Supplementary Fig. 1). Arrays of graphene r.f. transistors on a DLC substrate were fabricated using a conventional top-down approach. The schematic view of the graphene r.f. transistor is shown in Fig. 1a. Electron-beam lithography was used to define the channel, source and drain contacts, and the gate electrodes. Oxygen plasma etching was used to remove graphene outside the channel. The source and drain contacts consist of a thin Pd film covered by a thicker Au film. The top-gate dielectric film includes an electron-beam-evaporated Al layer, which is then oxidized and an additional layer of Al_2O_3 grown by atomic layer deposition (ALD)[21]. We note that all fabrication process steps involve standard top-down approaches that can be readily implemented in high-throughput production. Figure 1b shows a scanning electron microscope (SEM) image of a dual-channel graphene r.f. transistor with a ground-signal-ground coplanar pad design suitable for r.f. measurements. Figure 1d shows an SEM image of the well-aligned fine structure of a device with a gate length of 40 nm. The transmission electron microscopy (TEM) image in Fig. 1c further confirms the excellent alignment between the gate and the source/drain electrodes and the gate length of 40 nm, the shortest demonstrated so far. This nearly perfect alignment with a small un-gated region of less than 20 nm in our transistors is critical for achieving good device performance. The access region between gate and source/drain is nearly constant for all the devices, regardless of their gate length.

Figure 1e shows the output characteristics of two graphene devices, one with a gate length of 550 nm (left) and one with a gate length of 40 nm (right). The drain voltage sweeps from 0 V to 1.6 V for the 550-nm device and from 0 to 1 V for the 40-nm device. The gate voltage changes from −8 V to 0 V, from bottom trace to top trace. As shown in the insets, the Dirac point voltage obtained from the long-channel (that is, long-gate) device is around −7 V as a result of impurity charge doping, possibly induced during the transfer process. We attribute this effect to fixed impurity doping rather than trap charges because of the very weak hysteresis and the temperature-independent position of the Dirac point observed in these devices. The gate modulation of the short-channel (that is, short-gate) device is much weaker than that of the long channel one. This is mainly due to the more dominant role of the contact resistance in short channel devices. Unlike the case of Si metal-oxide-semiconductor field-effect transistors (MOSFETs), there is currently no proven way to reduce the contact resistance of graphene transistors. Another cause of the weak modulation of the 40-nm device is the 'short-channel effect', in which the electrostatic control efficiency of the top gate is adversely affected by the drain voltage. Although this effect is well-studied for conventional Si MOSFETs, it is not well understood in graphene transistors, and may be even more severe in these devices owing to the conical graphene band structure and the occurrence of Klein tunnelling.

[1] IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598, USA.

©2011 Macmillan Publishers Limited. All rights reserved

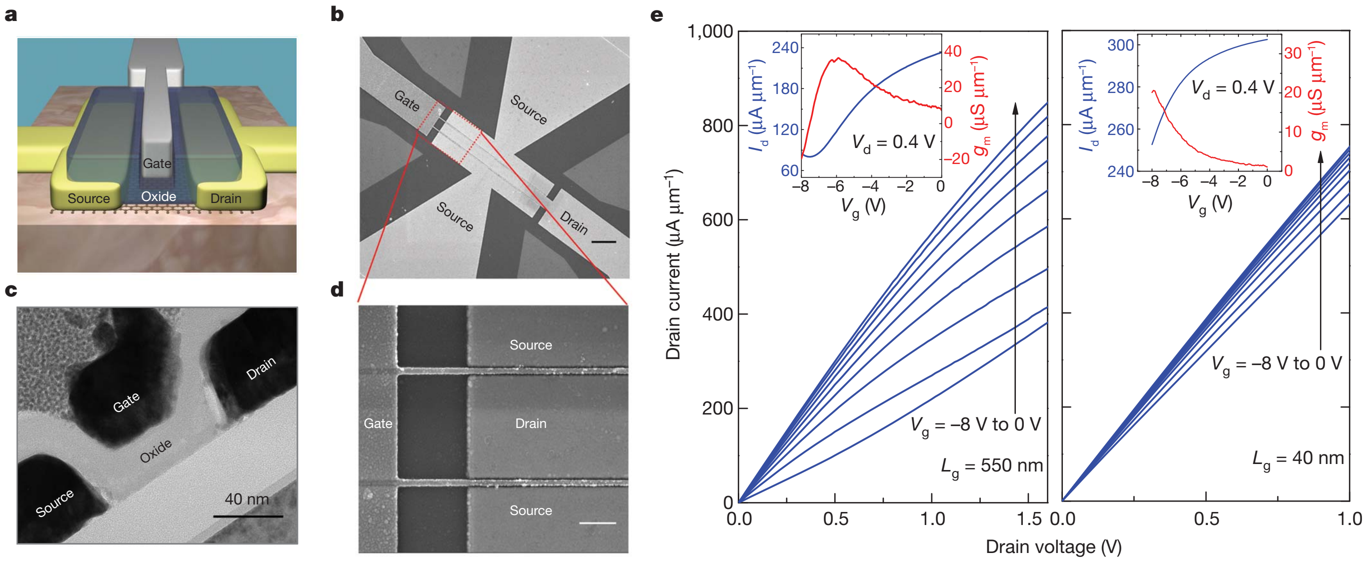


Figure 1 | Fabrication and output characteristics for graphene r.f. transistors. **a**, Schematic view of a top-gated graphene r.f. transistor on DLC substrate. **b**, SEM image of a typical top-gated dual-channel r.f. device. Scale bar, 3 μm. **c**, Cross-section TEM image of a graphene transistor with a gate length of 40 nm. Scale bar, 40 nm. **d**, SEM image of the 40-nm device. Scale bar, 400 nm. **e**, d.c. output characteristics of a 550-nm device (left) and a 40-nm device (right). Insets, transfer characteristics at drain–source voltage $V_{ds} = 0.4$ V.

High-frequency scattering parameters (S) of the graphene r.f. transistors were measured up to 30 GHz by an Agilent E8364C network analyser using standard ground-signal-ground probes (details of the measurement set-up and procedures are given in the Methods Summary). We used a de-embedding procedure that took account of parasitic effects (such as pad capacitance and wire resistance). This is achieved by measuring on-chip, inactive, 'open' and 'short' test devices; in the former, there was no graphene, and in the latter, gate, source and drain electrodes were all connected by metals. High fidelity in the de-embedding process was achieved by ensuring that the layouts of these open and short structures were strictly identical to that of the active device. The cut-off frequency (f_T), defined as the frequency at which the current gain becomes unity, is one of the most important figures-of-merit for evaluating the performance of r.f. devices. In Fig. 2a, the current gain, calculated from S parameters, is plotted against frequency f; a peak cut-off frequency f_T of 26 GHz is obtained for the 550-nm-long device at room temperature. To verify the value of f_T independently, Gummel's method[22,23] was adopted, and the same value of f_T was obtained, as shown in the inset of Fig. 2a. The current gain of devices with shorter gate lengths ($L_g = 140$ nm and 40 nm) is plotted in a similar fashion in Fig. 2b and c, respectively. A cut-off frequency of 70 GHz was obtained for the 140-nm transistor from both the intercept of the $1/f$ dependence and Gummel's method. An f_T as high as 155 GHz was obtained from the 40-nm-long device; this is the highest cut-off frequency yet achieved on CVD graphene, and 40 nm is also the smallest gate length reported so far. Although the direct current transconductance (g_m) suffers from the short-channel effect at this gate dimension, as discussed above, the overall r.f. performance benefits from the reduction of gate length.

In a well-behaved field-effect transistor (FET), the cut-off frequency can be related to g_m by the following equation[9]: $f_T = g_m/2\pi C_g$, where $C_g = \varepsilon_0\varepsilon_r W_g L_g/t_{ox}$. Here C_g is the gate capacitance, ε_0 is the dielectric constant of vacuum, ε_r is the relative dielectric constant of the gate dielectric, W_g is the channel width, L_g is the gate length and t_{ox} is the gate dielectric thickness. Therefore, the product $f_T L_g$ is expected to be linearly proportional to g_m for devices with the same gate dielectric and width, and the slope of a plot of $f_T\, L_g$ versus g_m is determined only by the physical thickness of the gate oxide. Such a plot of $f_T L_g$ against g_m for three different gate lengths is shown in Fig. 3a; it exhibits the expected linear dependence, with a slope independent of the gate

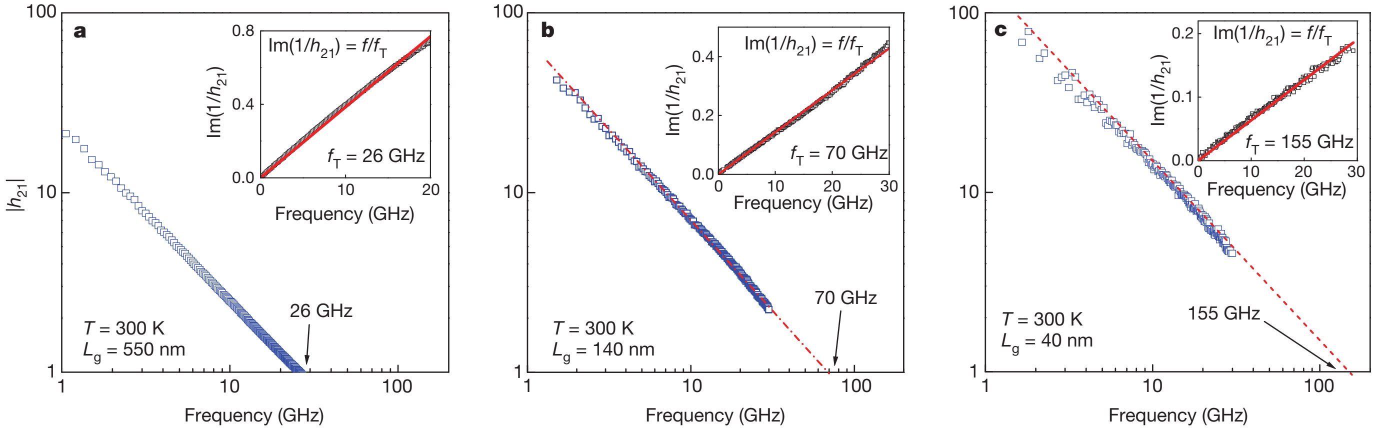


Figure 2 | Cut-off frequencies for three different devices at room temperature. Small-signal current gain $|h_{21}|$ versus frequency for devices with a gate length of 550 nm (**a**), 140 nm (**b**) and 40 nm (**c**) at room temperature. Intercepts give the cut-off frequency as 26 GHz, 70 GHz and 155 GHz, respectively. Insets, linear fitting using Gummel's method, showing extrapolated cut-off frequencies identical to the value obtained in the main panel for each device.

©2011 Macmillan Publishers Limited. All rights reserved

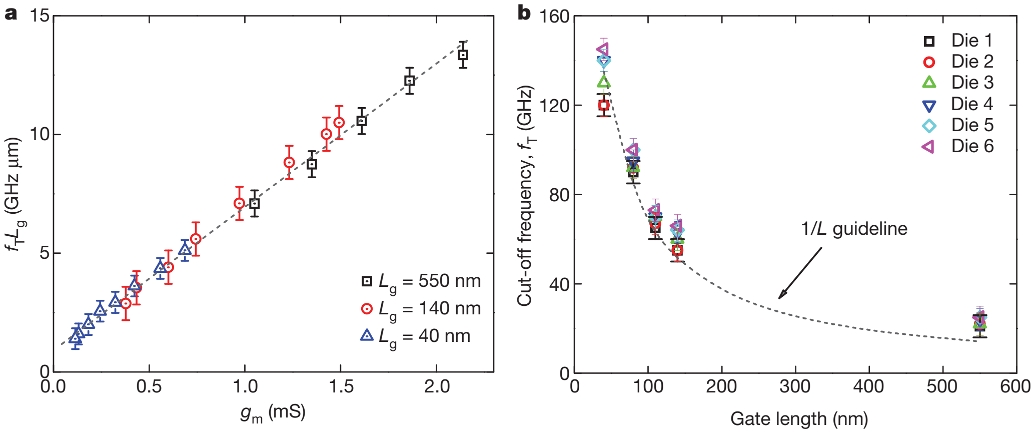


Figure 3 | Scaling behaviour of cut-off frequencies with gate length down to 40 nm. a, $f_T L_g$ versus direct current transconductance for three gate lengths: 550 nm (black squares), 140 nm (red circles) and 40 nm (blue triangles). The data from three different types of devices fall onto the same line, the slope of which corresponds to the unit area gate capacitance. This shows the uniformity of the graphene devices, the consistency of the measurements and the accuracy of the de-embedding approach. Data are shown as mean ± s.d., $n = 6$. **b**, Peak f_T as a function of gate length from 30 devices in 6 different dies located on the same wafer. Data are fitted well by the curve showing a $1/L_g$ dependence. Data are shown as mean ± s.d., $n = 3$.

lengths. This shows the uniformity of our devices across the whole wafer, and also demonstrates the measurement reliability. A high value of $f_T L_g = 13$ GHz µm is obtained for the 550-nm device; this is significantly higher than the value of 9 GHz µm for Si MOSFETs obtained from the International Technology Roadmap for Semiconductors (ITRS)[24], and is getting close to the best experimental results for Si MOSFETs.

To further test device uniformity and to examine device variation across the wafer, a systematic study of graphene transistors with five different gate lengths was performed; the results are shown in Fig. 3b. For each gate length, six devices from different dies (each die contains a complete set of devices) on the same wafer are measured and the peak cut-off frequencies are plotted. The performance variation is very small for all devices with the same gate length, as can be seen from Fig. 3a and b. Also, a $1/L_g$ dependence (the 'scaling trend') of the peak cut-off frequency is shown here, and is valid for devices with short gate lengths, even at the scaling limit of 40 nm. Previously, a $1/L_g^2$ dependence for f_T was observed for graphene FETs with long gate lengths where the transport is channel-resistance limited, partly owing to the severe mobility degradation associated with non-optimized gate dielectrics[8]. The $1/L_g$ scaling trend observed in our devices indicates that the transport is in a contact-limited regime, so that the electric field along the channel is dominated by the value of contact resistance at the source and drain and has little gate-length dependence. We note here that a similar $1/L_g$ dependence is usually observed for short-channel conventional Si and III–V FETs. This dependence is mainly due to the nearly-constant effective carrier velocity (obtained by reaching the saturation velocity of the material), which is seldom observed in current graphene devices. The long-channel (550-nm) device is in the region of transition from channel-limited transport to contact-limited transport. The f_T values obtained for this gate length are higher than the $1/L_g$ trend line; this is partly due to minimal short-channel effect for this relatively long gate.

In devices made from conventional semiconductors, the electrostatic potential profile will be controlled partly by the drain bias when the gate length is reduced. The resulting threshold voltage shift and drain-induced barrier lowering cause deterioration of the transistor switching. As in Si MOSFETs, controlling short-channel behaviour is of vital importance for graphene transistors. Moreover, occurrence of Klein tunnelling in graphene p–n junctions would make the short-channel effect worse[25–27]. Therefore, the transconductance is expected to decrease upon gate length scaling. The trade-off between performance and small device size will be the key factor in determining the scaling limit of graphene transistors. Nevertheless, as shown in Figs 2c, 3b and 4c, the 40-nm top-gated CVD graphene transistor on DLC is still well-behaved, with high r.f. performance. The result also shows the great potential for graphene transistors to be scaled even further, to a much smaller device size.

Many interesting studies of graphene physics, including investigations of the quantum Hall effect, have been performed at low temperatures: in contrast, no graphene r.f. transistors have been operated and studied below room temperature. A number of applications require cryogenic operation of r.f. devices, and how the graphene r.f. device performs at low temperatures is also scientifically important. Here we carried out the first study of graphene r.f. transistors down to liquid helium temperatures. Care was taken to ensure measurement accuracy; this included calibrating the system after the probe temperature and sample temperature had reached equilibrium. We found that—unlike many previous studies of graphene on SiO_2 substrates which showed strong temperature-dependent interface trap density and occupation—the r.f. performance of the graphene devices on DLC showed very little, if any, temperature dependence, being essentially unchanged at 4.3 K (Fig. 4a and b). Different gate biases were applied (from −8 V to 0 V) with a fixed drain bias of 1.6 V in a 550-nm-long device, where the current gain follows a $1/f$ dependence between 300 K and 4.3 K. This stability with temperature illustrates a significant advantage of the high quality DLC substrate, in which the trap density is very low. Figure 4c shows the current gain as a function of frequency for three devices with different gate lengths at 4.3 K; all exhibit a well-defined $1/f$ dependence, as at room temperature. As shown in the summary plot in Fig. 4d, the cut-off frequency shows little temperature dependence in the range from 300 K to 4.3 K. Unlike the carrier freeze-out effects typically observed in Si MOSFETs at cryogenic temperatures[28], the consistent temperature-independent results found here open up new opportunities for future graphene r.f. applications, such as ultra-low-noise or outer-space operations.

Besides the cut-off frequency, another important figure of merit for r.f. devices is the available power gain[28,29]; this is assessed using the maximum oscillation frequency (f_{MAX}), defined as the frequency at which the power gain is equal to one. Very low power gain has been achieved previously with graphene r.f. devices, and it was therefore seldom reported. The poor f_{MAX} of graphene devices usually results from the lack of clear current saturation and non-optimized gate structure. Here we show that despite the lack of clear saturation, we can achieve a high f_{MAX} of 20 GHz from the 550-nm device and 13 GHz from the 140-nm device (see Supplementary Fig. 4). It is noted that f_{MAX}, unlike f_T, is highly dependent on the design of the device and on the details of the interconnects, such as the gate metal thickness.

©2011 Macmillan Publishers Limited. All rights reserved

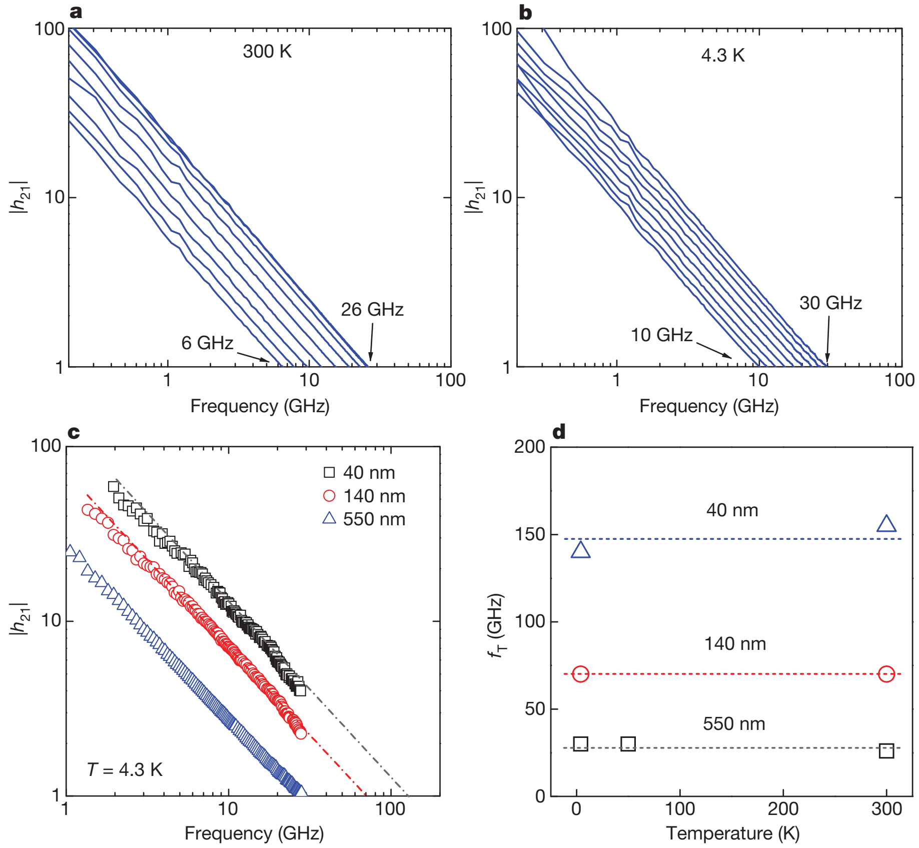


Figure 4 | Temperature dependence of cut-off frequency for different devices. **a**, **b**, Current gain as a function of frequency at 300 K (**a**) and 4.3 K (**b**). The gate length is 550 nm, with a V_{ds} of 1.6 V and with V_{gs} varying from −8 V to 0 V. **c**, Current gain versus frequency for three values of L_g (550 nm, 140 nm and 40 nm) at 4.3 K. The value of f_T is 28 GHz, 70 GHz and 140 GHz, respectively. **d**, Summary plot of the temperature dependence of f_T for three different devices; little temperature dependence was found.

Use of an optimized design, such as a mushroom-shaped gate to reduce the gate resistance, is expected to further improve f_{MAX}.

The r.f. performance of graphene is limited mainly by two factors: the substrate-limited carrier mobility and the contact resistance. Whereas the mobility dominates in long-channel devices, contact resistance becomes more critical as the gate length decreases. To further improve the r.f. performance of the graphene devices, efforts should be made to minimize the contact resistance, as short-channel devices are best suited to achieving ultimate device performance and high-density circuits. The contact resistance of graphene transistors is currently typically up to an order of magnitude higher than that of Si MOSFETs. Also, the short-channel effect can be mitigated by scaling down the thickness of the gate dielectric to achieve better electrostatic control by the gate.

METHODS SUMMARY

Top-gated graphene r.f. transistors were fabricated using graphene grown by CVD on copper[14], as follows. After evacuation of the CVD chamber, the Cu foil was heated to 875 °C in forming gas (H_2/Ar) and kept at this temperature for 30 min. After reduction, the Cu foil was exposed to ethylene at 975 °C for 10 min and then cooled. PMMA was spin-coated on top of the graphene layer that had formed on one side of the Cu foil. The Cu foil was then dissolved in 1 M iron chloride solution. The remaining graphene/PMMA layer was washed and transferred to the desired substrate. Subsequently, the PMMA was dissolved by treatment with hot acetone for one hour. The CVD graphene after transfer to DLC was characterized by Raman spectroscopy before device fabrication. DLC film was grown on an 8-inch Si substrate using cyclohexane (C_6H_{12}) with a vapour pressure of 1.8 p.s.i. in a CVD chamber. The flow rate was typically 25–40 cm^3 STP per min at 100 mtorr pressure. The DLC growth rate is 32 Å s^{-1} at 60 °C; this was followed by an anneal step at 400 °C for 4 h. The source/drain contact was 20 nm Pd/30 nm Au deposited by electron-beam evaporation. The gate oxide was formed by an oxidized Al layer deposited by electron-beam evaporation, followed by the deposition of 15 nm ALD Al_2O_3 film. The direct current and r.f. characterizations were carried out in a probe station under $<10^{-6}$ torr using an Agilent parameter analyser B1500, and a E8364C network analyser. The system was calibrated using a short-open-load-through method. On-chip open and short structures with the exact design of the devices were used to de-embed parasitic effects, such as pad capacitance and interconnection resistance. The low-temperature measurements were performed using the same approach, and system calibration was done for each temperature. On-chip de-embedding, using standard open-short structures, was also done at each temperature.

Received 3 January; accepted 4 March 2011.

1. Novoselov, K. S. *et al.* Two-dimensional gas of massless Dirac fermions in graphene. *Nature* **438,** 197–200 (2005).
2. Zhang, Y. B., Tan, Y. W., Stormer, H. L. & Kim, P. Experimental observation of the quantum Hall effect and Berry's phase in graphene. *Nature* **438,** 201–204 (2005).
3. Berger, C. *et al.* Electronic confinement and coherence in patterned epitaxial graphene. *Science* **312,** 1191–1196 (2006).
4. Avouris, P. Graphene: electronic and photonic properties and devices. *Nano Lett.* **10,** 4285–4294 (2010).
5. Lemme, M. C., Echtermeyer, T. J., Baus, M. & Kurz, H. A graphene field-effect device. *IEEE Electron Device Lett.* **28,** 282–284 (2007).
6. Lin, Y.-M. *et al.* Operation of graphene transistors at gigahertz frequencies. *Nano Lett.* **9,** 422–426 (2009).
7. Meric, I., Baklitskaya, N., Kim, P. & Shepard, K. L. RF performance of top-gated, zero-bandgap graphene field-effect transistors. *Tech. Dig. IEDM* 1–4 doi:10.1109/IEDM.2008.4796738 (2008).
8. Lin, Y.-M. *et al.* Development of graphene FETs for high frequency electronics. *Tech. Dig. IEDM* 237–240 doi:10.1109/IEDM.2009.5424378 (2009).
9. Lin, Y.-M. *et al.* 100 GHz transistors from wafer-scale epitaxial graphene. *Science* **327,** 662 (2010).
10. Moon, J. S. *et al.* Epitaxial-graphene RF field-effect transistors on Si-face 6H-SiC substrates. *IEEE Electron Device Lett.* **30,** 650–652 (2009).
11. Liao, L. *et al.* High-speed graphene transistors with a self-aligned nanowire gate. *Nature* **467,** 305–308 (2010).
12. Lee, J. *et al.* RF performance of pre-patterned locally-embedded-back-gate graphene device. *Tech. Dig. IEDM* 568–571 doi:10.1109/IEDM.2010.5703422 (2010).

©2011 Macmillan Publishers Limited. All rights reserved

13. Schwierz, F. Graphene transistors. *Nature Nanotechnol.* **5,** 487–496 (2010).
14. Li, X. S. *et al.* Large-area synthesis of high-quality and uniform graphene films on copper foils. *Science* **324,** 1312–1314 (2009).
15. Han, M. Y., Özyilmaz, B., Zhang, Y. B. & Kim, P. Energy band-gap engineering of graphene nanoribbons. *Phys. Rev. Lett.* **98,** 206805 (2007).
16. Wang, X. *et al.* Room-temperature all-semiconducting sub-10-nm graphene nanoribbon field-effect transistors. *Phys. Rev. Lett.* **100,** 206803 (2008).
17. Chen, Z., Lin, Y.-M., Rooks, M. J. & Avouris, Ph Graphene nano-ribbon electronics. *Physica E* **40,** 228–232 (2007).
18. Li, X., Wang, X., Zhang, L., Lee, S. & Dai, H. Chemically derived, ultrasmooth graphene nanoribbon semiconductors. *Science* **319,** 1229–1232 (2008).
19. Xia, F., Farmer, D. B., Lin, Y.-M. & Avouris, Ph Graphene field-effect transistors with high on/off current ratio and large transport band gap at room temperature. *Nano Lett.* **10,** 715–718 (2010).
20. Robertson, J. Diamond-like amorphous carbon. *Mater. Sci. Eng. Rep.* **37,** 129–281 (2002).
21. Lee, B. K. *et al.* Conformal Al_2O_3 dielectric layer deposited by atomic layer deposition for graphene-based nanoelectronics. *Appl. Phys. Lett.* **92,** 203102 (2008).
22. Kim, D. H. & del Alamo, J. A. 30-nm InAs pseudomorphic HEMTs on an InP substrate with a current-gain cutoff frequency of 628 GHz. *IEEE Electron Device Lett.* **29,** 830–833 (2008).
23. Gummel, H. K. On the definition of the cutoff frequency f_T. *Proc. IEEE* **57,** 2159 (1969).
24. Arden, W. *et al.* (eds) ITRS 2009 edition. ⟨http://www.itrs.net/Links/2009ITRS/Home2009.htm⟩ (21 June 2010).
25. Katsnelson, M. I., Novoselov, K. S. & Geim, A. K. Chiral tunnelling and the Klein paradox in graphene. *Nature Phys.* **2,** 620–625 (2006).
26. Huard, B. *et al.* Transport measurements across a tunable potential barrier in graphene. *Phys. Rev. Lett.* **98,** 236803 (2007).
27. Young, A. F. & Kim, P. Quantum interference and Klein tunnelling in graphene heterojunctions. *Nature Phys.* **5,** 222–226 (2009).
28. Sze, S. M. & Ng, K. K. *Physics of Semiconductor Devices* 3rd edn (Wiley-Interscience, 2006).
29. Lee, T. H. *The Design of CMOS Radio-Frequency Integrated Circuits* (Cambridge Univ. Press, 2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Graham, B. Ek, J. Bucchignano, C. V. Jahnes and S. Han for technical assistance, and C. Y. Sung, V. Perebeinos, A. Valdes-Garcia, C. Dimitrakopoulos, W. Zhu and H.-Y. Chiu for discussions. This work was supported in part by DARPA through the CERA programme (contract FA8650-08-C-7838). The views, opinions and findings contained in this Letter are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of DARPA or the US Department of Defense.

Author Contributions Y.W., Y.-m.L. and P.A. designed the experiment, and Y.W. performed device fabrication, electrical characterization and data analysis. Y.-m.L. and K.A.J. contributed to the r.f. characterization. A.A.B. performed graphene synthesis, and F.X. helped to prepare the DLC substrate. Y.-m.L and D.B.F. contributed to device fabrication. Y.Z. performed TEM imaging. Y.W. wrote the Letter, and Y.-m.L. and P.A. discussed and commented on the manuscript. All authors provided feedback.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to Y.-m.L. (yming@us.ibm.com) and P.A. (avouris@us.ibm.com).

©2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09926

Evidence for mechanical coupling and strong Indian lower crust beneath southern Tibet

Alex Copley[1]†, Jean-Philippe Avouac[1] & Brian P. Wernicke[1]

How surface deformation within mountain ranges relates to tectonic processes at depth is not well understood. The upper crust of the Tibetan Plateau is generally thought to be poorly coupled to the underthrusting Indian crust because of an intervening low-viscosity channel[1]. Here, however, we show that the contrast in tectonic regime between primarily strike-slip faulting in northern Tibet and dominantly normal faulting in southern Tibet requires mechanical coupling between the upper crust of southern Tibet and the underthrusting Indian crust. Such coupling is inconsistent with the presence of active 'channel flow' beneath southern Tibet, and suggests that the Indian crust retains its strength as it underthrusts the plateau. These results shed new light on the debates regarding the mechanical properties of the continental lithosphere[2–4], and the deformation of Tibet[1,5–10].

The processes governing continental deformation, and the formation of mountain ranges and plateaus, are hotly debated[2,3,8,10]. Because it is the largest mountain range on the Earth, and has been formed by processes that are still active, the Tibetan Plateau has been central in this debate and has inspired a wide range of tectonic models. In 1924 Argand[11] proposed that Indian crust underthrusts most of Tibet, and that the resulting doubling of crustal thickness is responsible for the high elevation of the plateau; a view which has to some extent been confirmed by recent geophysical observations that suggest that the Indian crust underlies the southern half of the plateau[12]. This view is also consistent with the large amount of underthrusting implied by kinematic models of the orogen derived from structural geology[13] and the metamorphic and exhumation history of the range[14].

However, how the underthrusting of India influences the tectonics of Tibet is unclear. High temperatures (over 600 °C) must exist in the deep crust of Tibet, as suggested by heatflow measurements[15] and thermokinematic models[14]. Various geophysical observations[16] have been interpreted as evidence for a 'channel' of weak, possibly partially molten, middle crust beneath southern Tibet. The middle crust of Tibet may therefore have a low enough viscosity to result in mechanical decoupling between the Tibetan upper crust and the underthrusting Indian lithosphere. A popular extension of this view is that the middle crust might actually be extruded from below the high topography, both southwards towards the Himalaya[1,17] and eastwards towards southeast Asia[10]. On the other hand, some authors have argued that the whole Tibetan lithosphere might actually be deforming as a coherent unit, with little depth variation of strain[7].

The deformation of Tibet arises from the forces driving the India–Asia collision: essentially the buoyancy of the Indian ridge and the sinking of subducting slabs beneath southeast Asia[18]. In addition, forces are induced within the plateau and bounding mountain ranges by the lateral variations of crustal thickness[7,8]. The Tibetan crust is approximately 75 km thick, about twice the thickness of the relatively undeforming continental crust in the surrounding areas[19]. This contrast is certainly one key factor in determining the state of stress within the plateau, as demonstrated by the correlation between elevation and tectonic regime[6]: thrust faulting is dominant at low elevations around the edge of the mountain range, whereas the high interior of the plateau deforms by a combination of normal and strike-slip faulting[20–22] (Fig. 1). Mantle dynamics could also play a part, but the

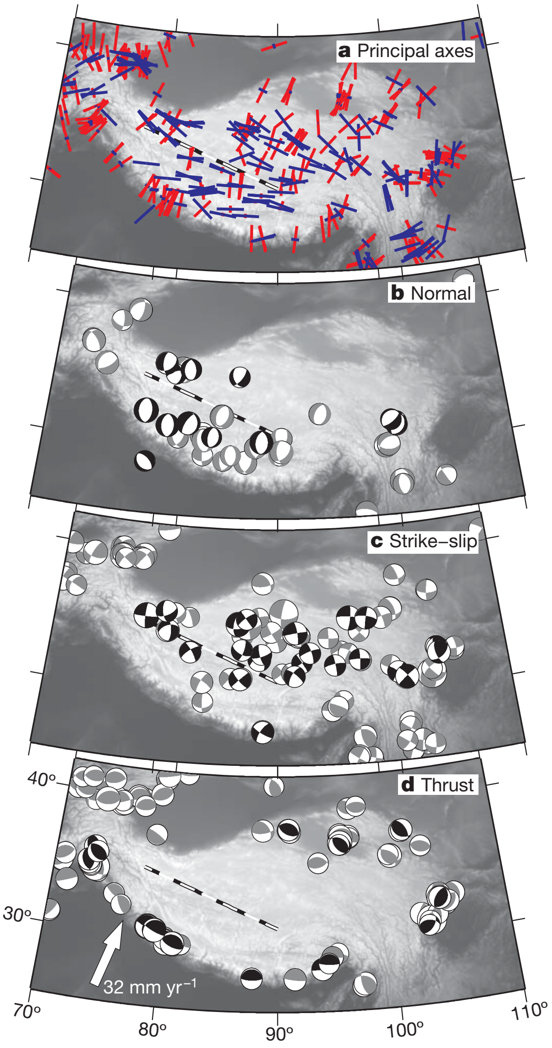


Figure 1 | Tectonic regime within and around the Tibetan Plateau. **a**, Principal axes of the horizontal components of the earthquake moment tensors, normalized to the length of the largest axis (red is compression, blue is extension). **b**, **c** and **d**, Focal mechanisms of upper crustal (depth less than 50 km) earthquakes of moment magnitude exceeding 5.5, subdivided on the basis of rake. Black focal mechanisms are from the studies listed in the Supplementary Information; grey focal mechanisms are well-constrained CMT solutions (http://www.globalcmt.org/; over 50% double couple; ref. 30). **d** also shows the India–Asia convergence velocity[23]. The dashed line in the central plateau on each panel shows the estimated location of the northern limit of underthrust Indian lithosphere[12,19].

[1]Tectonics Observatory, Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA. †Present address: Bullard Labs, Department of Earth Sciences, University of Cambridge, CB3 0EZ Cambridge, UK.

©2011 Macmillan Publishers Limited. All rights reserved

hypothesis that thickened mantle lithosphere has been convectively removed from beneath the range[6] can now be ruled out because of the observation that Tibet is still underlain by a continuous mantle lid visible to surface wave tomography[19].

Some previous attempts at modelling Tibetan tectonics as a result of crustal buoyancy, and of north–south compression induced by the collision, have yielded good agreement with the distribution of present-day strain around Tibet[8]. Such studies reproduce the contrast between thrust faulting around the edge of the plateau and east–west extension within the range, but a close look at the active deformation within the plateau indicates a clear contrast between southern and northern Tibet that is not explained by existing models. Earthquake focal mechanisms (Fig. 1) and mapped active faults show that the deformation of southern Tibet is dominated by east–west extension across north–south-trending rifts[20], whereas northern Tibet is characterized by conjugate strike-slip faulting (with some minor normal faulting also occurring at fault bends and junctions[21]). It should be noted that the north–south shortening observed in Global Positioning System (GPS) data within the southern plateau represents recoverable elastic strain build-up around the thrust faults beneath the Himalayas[23], and not the permanent deformation with which we are concerned here. Any shortening within the southern plateau that cannot be explained by elastic strain around the Himalayan thrust faults is lower in magnitude than is resolvable with the currently available GPS data, and so is minor compared with the east–west extension that is geodetically visible and is accommodated by the observed normal faulting (see Supplementary Information).

The contrast between north and south Tibet is not likely to be due to lateral variations in topographically induced stresses, given the uniform elevation of the plateau. We observe that the change in tectonic regime, which occurs at the Karakoram–Jiali fault zone that runs between the eastern and western Himalayan syntaxes[21], coincides approximately with the proposed location of the northern edge of the underthrust Indian crust and upper mantle[12,19,24]. We therefore investigate whether mechanical coupling between the Tibetan upper crust and underthrust Indian crust could actually explain the contrast in present-day tectonic regime between southern and northern Tibet. Such an idea is plausible because the underthrusting Indian crust will exert considerable northward-directed shear stresses upon the overlying material, which are not likely to be present in northern Tibet, thereby leading to a fundamental difference in stress state between the two regions. To test this hypothesis we have modelled the active deformation of Tibet, resulting from approximately north–south compression induced by the collision, and lateral variations in crustal thickness. We have assumed either coupling to, or decoupling from, the underthrusting Indian crust, which is modelled as either rigid or viscously deforming.

Following many previous investigations of continental tectonics, we assume that the crust obeys a viscous rheology[5,6,25]. We acknowledge that this modelling cannot reproduce the details of surface tectonics, which are locally characterized by deformation on discrete faults. However, the model is appropriate for estimating how large-scale lateral variations of tectonic regime within Tibet depend upon the boundary conditions around the edge of the plateau (which we impose on the basis of GPS measurements), and those at the base of the deforming crust (which is the effect we study here). A previous study analysed the decoupling effect of a weak middle crust in two dimensions[9], but did not address the effect of such a weak horizon on the spatial variations of tectonic regime within Tibet. This question, which we pursue here, requires a three-dimensional model. We therefore use the approach of Copley[25], assuming a two-layered viscosity structure based upon previous studies[5,25] (see Methods).

We compare three numerical experiments. In experiment A (Fig. 2a), the lower 20 km of the underthrusting crust beneath the southern half of the plateau is assumed to be rigid. In the southern plateau the surface motions are accommodated by the shearing of the upper crust over this rigid lower crust, leading to significant shear stresses on horizontal planes. Where the topography slopes steeply on the southern margin of the range, topographically induced stresses dominate the deformation and lead to arc-normal compression. East–west extension of the upper crust within the southern plateau is caused by the combination of the shear stresses on horizontal planes, the topographically induced stresses that are transmitted to the interior of the range, and the approximately north–south compression imposed by the applied motions of the bounding plates. In models with topographic forces and convergence across the range, the effect of the horizontal shear stresses related to the underthrust rigid lower crust is to make the southern plateau interior move more slowly southwards than it otherwise would (equivalent to the overlying crust feeling a pull northwards by the underthrust crust). The resulting north–south extensional stresses between this region and the southern margin balance the compression resulting from the plate convergence. The relative contributions of these causes of deformation are shown in the Supplementary Information. In this model, the weak middle crust of the southern plateau does not flow southwards as a high-velocity channel, but rather acts as a horizontal simple shear zone, transmitting to the upper crust the shear that is induced by the relative motion between the surface and

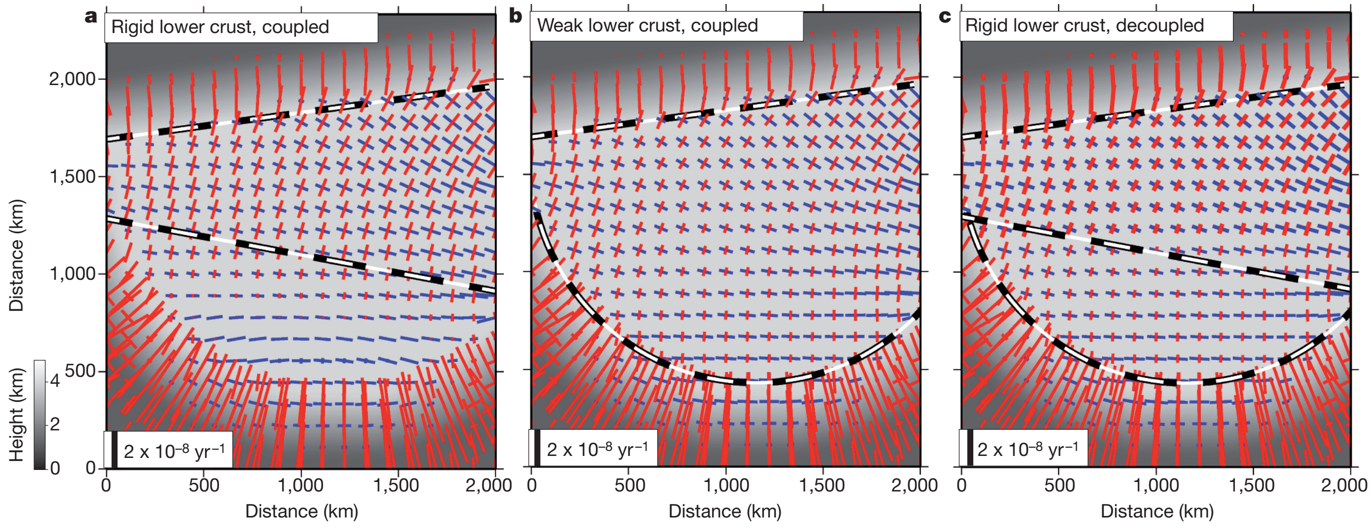


Figure 2 | Modelled principal axes of the horizontal strain-rate tensor at the surface. Red bars represent compression, and blue bars extension. Red and blue crosses (with bars of equal length) indicate strike–slip deformation. North of the northernmost dashed line, the lower 35 km of the crust is given the velocity of Tarim relative to India. For **a** and **b**, south of the southernmost dashed line the lower 20 km of the crust is forced to have zero velocity. **c** is the same as **a**, except that between the two southernmost dashed lines a horizontal decoupling horizon is inserted above the rigid lower crust. Background shading represents elevation. See Supplementary Information for the modelled velocities. Scale bars are strain rate, 2×10^{-8} yr^{-1}.

©2011 Macmillan Publishers Limited. All rights reserved

the underthrusting lower crust. The northern plateau is characterized by strike–slip deformation (assuming that the tectonic regime is related to the stress tensor according to Anderson's theory of faulting). The tectonic style differs from the southern plateau because in this northern region the shear stresses on horizontal planes are negligible.

In experiment B (Fig. 2b) we impose the condition that the rigid lower crust extends only a short distance beneath the southern margin of the plateau. The interior of the range in this case is everywhere characterized by strike–slip deformation. This is because shear stresses on horizontal planes are negligible throughout the interior of the range. In the southern plateau in experiment A, it was these shear stresses that had the effect of counteracting the compression imposed upon the range by the motions of the bounding plates, allowing pure east–west extension to occur.

Experiment C (Fig. 2c) is similar to experiment A, but with the addition of a decoupling horizon above the rigid lower crust, where shear stresses on horizontal planes are forced to be zero. This model behaves very similarly to that in experiment B, because they share the characteristic that no significant shear stresses on horizontal planes are present in the middle and upper crust.

Comparison between the results of our numerical experiments (Fig. 2) and the heterogeneous active deformation within the Tibetan Plateau (Fig. 1) suggests that at the present day the Indian lower crust acts in a rigid manner where it underlies southern Tibet, and that the surface is mechanically coupled to the lower crust in this region. The deformation in the northern plateau is similar (except for slightly different strain rates) in all three numerical experiments, showing the tectonics in this region to be relatively insensitive to the rheology of the underthrust Indian crust beneath the southern plateau. For the lower crust to act rigidly in numerical experiment A requires a viscosity of more than 5×10^{23} Pa s. Such a high viscosity at lower crustal temperatures would require an anhydrous rheology, such as metastable granulite[3]. Evidence of a strong rheology for the Indian lower crust, and an absence of large-scale granulite-to-eclogite transformation, have independently been inferred from the modelling of gravity anomalies across the Himalaya[26]. Mechanical coupling between the surface and the rigid lower crust implies an absence of low-viscosity decoupling horizons within the crust, and is therefore inconsistent with 'channel flow' models of present-day tectonics in southern Tibet.

METHODS SUMMARY

The model geometry and topography approximate what is currently seen in the Tibetan Plateau, and deformation is driven by velocity boundary conditions and topographically induced stresses. We have used a crustal thickness of 40 km under the lowlands in the north and south of the model, 75 km in the region of underthrust Indian lithosphere, and 65 km in the northern plateau[27]. The crustal thickness is tapered between the values used in the mountains and the lowlands in proportion to the surface topography. The perpendicular component of the velocity on the eastern and western boundaries is approximated and interpolated from GPS velocities[28], and no constraints are imposed on the component parallel to the boundary. The model is constructed in a reference frame attached to the lowlands in the southern part of the model domain, which represent northern India. For simplicity, a newtonian rheology is used throughout. The viscosity of the upper 15 km of the crust is 10^{22} Pa s (ref. 25), and that of the lower crust is 10^{20} Pa s (ref. 5). The viscosity is vertically tapered for 5 km either side of the contrast. Northern Tibet is underthrust by the Tarim basin for about 200 km (ref. 29). As in southern Tibet, we model this as a region of rigid lower crust (the Tarim basin, like India, is underlain by Precambrian basement), which is given the velocity of the central Tarim basin relative to India[28]. We assume that the vertical normal stress at the base of the model balances the mass of the overlying rock. We also impose zero shear stress on the base of the model, because some models of southeastern Tibet[25] suggested that the hot and hydrated mantle in the region was too weak to provide a rigid lower boundary to deformation within the crust.

Received 24 September 2010; accepted 10 February 2011.

1. Beaumont, C., Jamieson, R. A., Nguyen, M. H. & Lee, B. Himalayan tectonics explained by extrusion of a low-viscosity crustal channel coupled to focused surface denudation. *Nature* **414,** 738–742 (2001).
2. Watts, A. B. & Burov, E. B. Lithospheric strength and its relationship to the elastic and seismogenic layer thickness. *Earth Planet. Sci. Lett.* **213,** 113–131, doi:10.1016/S0012–821x(03)00289–9 (2003).
3. Jackson, J., Mckenzie, D., Priestley, K. & Emmerson, B. New views on the structure and rheology of the lithosphere. *J. Geol. Soc. Lond.* **165,** 453–465 (2008).
4. Hetenyi, G. *et al.* Density distribution of the India plate beneath the Tibetan plateau: Geophysical and petrological constraints on the kinetics of lower-crustal eclogitization. *Earth Planet. Sci. Lett.* **264,** 226–244, doi:10.1016/j.epst.2007.09.036 (2007).
5. Copley, A. & McKenzie, D. Models of crustal flow in the India-Asia collision zone. *Geophys. J. Int.* **169,** 683–698 (2007).
6. England, P. & Houseman, G. Extension during continental convergence, with application to the Tibetan plateau. *J. Geophys. Res.* **94,** 17561–17579 (1989).
7. England, P. & Molnar, P. Active deformation of Asia: from kinematics to dynamics. *Science* **278,** 647–650 (1997).
8. Flesch, L. M., Haines, A. J. & Holt, W. E. Dynamics of the India-Eurasia collision zone. *J. Geophys. Res.* **106,** 16435–16460 (2001).
9. Bendick, R. & Flesch, L. M. Reconciling lithospheric deformation and lower crustal flow beneath central Tibet. *Geology* **35,** 895–898 (2007).
10. Clark, M. K. & Royden, L. H. Topographic ooze: building the eastern margin of Tibet by lower crustal flow. *Geology* **28,** 703–706 (2000).
11. Argand, E. La tectonique de l'Asie. *Proc. 13th Int. Geological Congr.* **7,** 170–372 (1924).
12. Nabelek, J. *et al.* Underplating in the Himalaya-Tibet collision zone revealed by the Hi-CLIMB experiment. *Science* **325,** 1371–1374 (2009).
13. DeCelles, P. G., Robinson, D. M. & Zandt, G. Implications of shortening in the Himalayan fold-thrust belt for uplift of the Tibetan Plateau. *Tectonics* **21,** doi:10.1029/2001tc001322 (2002).
14. Bollinger, L., Henry, P. & Avouac, J. P. Mountain building in the Nepal Himalaya: thermal and kinematic model. *Earth Planet. Sci. Lett.* **244,** 58–71, doi:10.1016/j.epsl.2006.01.045 (2006).
15. Francheteau, J. *et al.* High heat-flow in southern Tibet. *Nature* **307,** 32–36 (1984).
16. Nelson, K. D. *et al.* Partially molten middle crust beneath southern Tibet: synthesis of project INDEPTH results. *Science* **274,** 1684–1688 (1996).
17. Grujic, D., Hollister, L. S. & Parrish, R. R. Himalayan metamorphic sequence as an orogenic channel: insight from Bhutan. *Earth Planet. Sci. Lett.* **198,** 177–191 (2002).
18. Copley, A., Avouac, J. P. & Royer, J. Y. India-Asia collision and the Cenozoic slowdown of the Indian plate: implications for the forces driving plate motions. *J. Geophys. Res.* **115,** doi:10.1029/2009jb006634 (2010).
19. Priestley, K., Jackson, J. & McKenzie, D. Lithospheric structure and deep earthquakes beneath India, the Himalaya and southern Tibet. *Geophys. J. Int.* **172,** 345–362 (2008).
20. Armijo, R., Tapponnier, P., Mercier, J. L. & Han, T. L. Quaternary extension in southern Tibet—field observations and tectonic implications. *J. Geophys. Res.* **91,** 13803–13872 (1986).
21. Taylor, M., Yin, A., Ryerson, F. J., Kapp, P. & Ding, L. Conjugate strike-slip faulting along the Bangong-Nujiang suture zone accommodates coeval east-west extension and north-south shortening in the interior of the Tibetan Plateau. *Tectonics* **22,** doi:10.1029/2002tc001361 (2003).
22. Tapponnier, P. & Molnar, P. Active faulting and tectonics in China. *J. Geophys. Res.* **82,** 2905 (1977).
23. Bettinelli, P. *et al.* Plate motion of India and interseismic strain in the Nepal Himalaya from GPS and DORIS measurements. *J. Geodesy* **80,** 567–589, doi:10.1007/s00190–006–0030–3 (2006).
24. Huang, W. C. *et al.* Seismic polarization anisotropy beneath the central Tibetan Plateau. *J. Geophys. Res.* **105,** 27979–27989 (2000).
25. Copley, A. Kinematics and dynamics of the southeastern margin of the Tibetan plateau. *Geophys. J. Int.* **174,** 1081–1100 (2008).
26. Cattin, R. *et al.* Gravity anomalies, crustal structure and thermo-mechanical support of the Himalaya of central Nepal. *Geophys. J. Int.* **147,** 381–392 (2001).
27. Tseng, T. L., Chen, W. P. & Nowack, R. L. Northward thinning of Tibetan crust revealed by virtual seismic profiles. *Geophys. Res. Lett.* **36,** doi:10.1029/2009gl040457 (2009).
28. Zhang, P. Z. *et al.* Continuous deformation of the Tibetan Plateau from global positioning system data. *Geology* **32,** 809–812 (2004).
29. Wittlinger, G. *et al.* Teleseismic imaging of subducting lithosphere and Moho offsets beneath western Tibet. *Earth Planet. Sci. Lett.* **221,** 117–130 (2004).
30. Jackson, J., Priestley, K., Allen, M. & Berberian, M. Active tectonics of the South Caspian basin. *Geophys. J. Int.* **148,** 214–245 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the Gordon and Betty Moore Foundation for support through the Caltech Tectonics Observatory, and Pembroke College in the University of Cambridge for financial support to A.C. This work benefited from a review by L. Flesch. This is Caltech Tectonics Observatory contribution number 145.

Author Contributions A.C. performed the calculations, A.C., J.-P.A. and B.W. discussed the results, and A.C. and J.-P.A. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.C. (acc41@cam.ac.uk).

©2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09927

Low strength of deep San Andreas fault gouge from SAFOD core

David A. Lockner[1], Carolyn Morrow[1], Diane Moore[1] & Stephen Hickman[1]

The San Andreas fault accommodates 28–34 mm yr^{-1} of right lateral motion of the Pacific crustal plate northwestward past the North American plate. In California, the fault is composed of two distinct locked segments that have produced great earthquakes in historical times, separated by a 150-km-long creeping zone. The San Andreas Fault Observatory at Depth (SAFOD) is a scientific borehole located northwest of Parkfield, California, near the southern end of the creeping zone. Core was recovered from across the actively deforming San Andreas fault at a vertical depth of 2.7 km (ref. 1). Here we report laboratory strength measurements of these fault core materials at *in situ* conditions, demonstrating that at this locality and this depth the San Andreas fault is profoundly weak (coefficient of friction, 0.15) owing to the presence of the smectite clay mineral saponite, which is one of the weakest phyllosilicates known. This Mg-rich clay is the low-temperature product of metasomatic reactions between the quartzofeldspathic wall rocks and serpentinite blocks in the fault[2,3]. These findings provide strong evidence that deformation of the mechanically unusual creeping portions of the San Andreas fault system is controlled by the presence of weak minerals rather than by high fluid pressure or other proposed mechanisms[1]. The combination of these measurements of fault core strength with borehole observations[1,4,5] yields a self-consistent picture of the stress state of the San Andreas fault at the SAFOD site, in which the fault is intrinsically weak in an otherwise strong crust.

SAFOD is a deep scientific borehole that penetrates the San Andreas fault (SAF) at a vertical depth of approximately 2.7 km and is the deepest land-based scientific drilling project to cross a plate-bounding fault[1,6,7] (see http://www.earthscope.org for additional information). During phase 2 drilling in 2005, the basic structure of the SAF was determined (Fig. 1) using borehole logging data[1] and supplementary laboratory studies of the drilling cuttings[8,9]. At 2.7 km depth, the damage zone associated with the fault is approximately 200 m wide, and two actively creeping strands were identified within it by accumulated deformation of the steel casing in the main borehole[1]. These two active shear zones, referred to as the southwest deforming zone (SDZ) and the central deforming zone (CDZ), were primary targets of the phase 3 multilateral core drilling operation in 2007. Approximately 31 m of core were recovered from across the SDZ, CDZ and adjoining damage-zone rocks, including 1.6 m and 2.6 m of highly foliated, incohesive fault gouge associated with the SDZ and CDZ, respectively.

We have completed frictional strength measurements on 25 core samples that span the important lithologic units. Of these, 17 are detrital sedimentary rocks, ranging from fine-grained sandstones to mudstones; representative X-ray diffraction (XRD) patterns are presented in Supplementary Fig. 2. The SDZ and CDZ are represented by four samples apiece. In marked contrast to the adjoining rocks, both foliated gouge zones consist of porphyroclasts of serpentinite and sedimentary rock dispersed in a matrix of Mg-rich clays[10] (Supplementary Fig. 3). XRD patterns of the CDZ were dominated by saponite (estimated to be greater than 60% from petrographic analysis) with some quartz and calcite. The SDZ gouge was composed primarily of saponite+corrensite with some quartz and feldspars (corrensite is a regularly interlayered chlorite-saponite clay). The porphyroclasts are also partly altered to Mg-rich clays[3]. The two gouge zones are interpreted to be the product of shearing-enhanced metasomatic reactions between serpentinite, tectonically entrained within the fault, and adjoining sedimentary rocks[2,3].

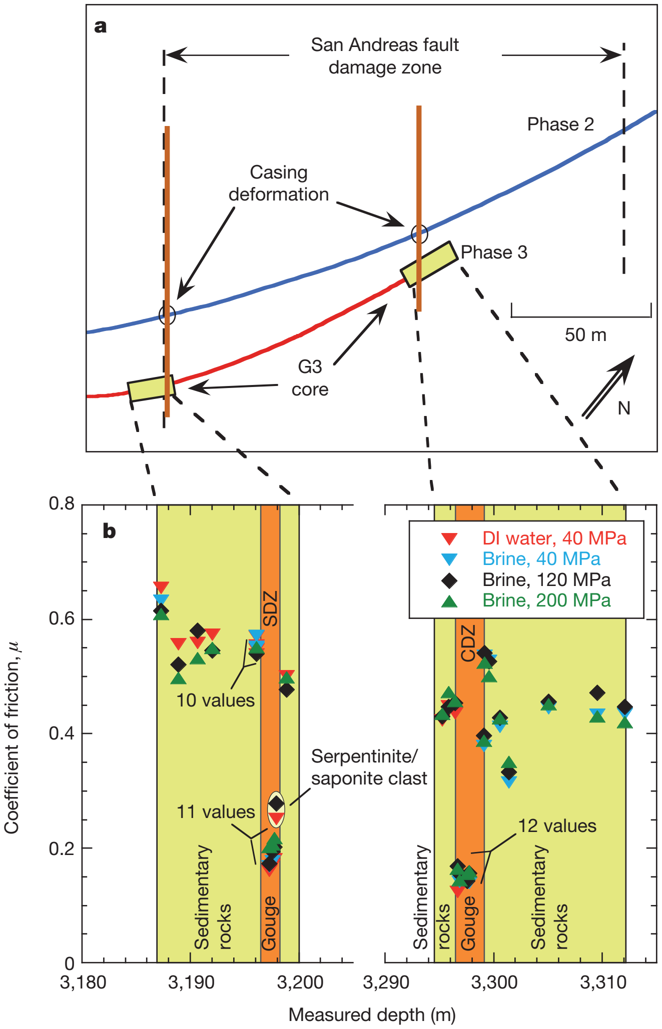


Figure 1 | Location and strengths of SAFOD core samples. a, Map view of SAF damage zone and SAFOD boreholes from phase 2 (blue) (indicating actively deforming casing) and phase 3 Hole G (red) with location of recovered core (yellow) at approximately 2.7 km vertical depth. Active deformation zones are shown in orange. **b**, Frictional strength of core samples plotted versus measured depth along Hole G (at sliding rate $V = 1.15\ \mu\mathrm{m\ s}^{-1}$). Active fault traces SDZ (3,196.4–3,198.1 m measured depth) and CDZ (3,296.6–3,299.1 m) have notably low strength. A few samples were tested with deionized (DI) water. Extrapolation to SAF plate rate reduces shear zone strength to $\mu \approx 0.15$.

[1]US Geological Survey, Menlo Park, California 94025, USA.

©2011 Macmillan Publishers Limited. All rights reserved

Sample strength is reported as coefficient of friction $\mu = \tau/\overline{\sigma_{\rm n}}$, where τ and $\overline{\sigma_{\rm n}}$ are respectively shear stress and effective normal stress on the test faults; we estimate *in situ* $\overline{\sigma_n}$ to be ~122 MPa (see below). Here, $\overline{\sigma_{\rm n}} = \sigma_{\rm n} - p$, where $\sigma_{\rm n}$ is normal stress and p is pore pressure. Representative strength tests are plotted in Fig. 2. Frictional strength was compiled from all deformation tests at 9.8–10.4 mm fault-parallel slip and sliding rate $V = 1.15\ \mu{\rm m\ s}^{-1}$. As shown in Fig. 2, nearly all time- and slip-dependent strengthening had ended by 10 mm slip, so that residual strength is reasonably represented by this value. Residual strength refers to the stable strength of the test sample once fully developed shear flow is established. Because control parameters—including $\sigma_{\rm n}$, p, V and pore fluid composition—are duplicated in the tests, variations in μ are attributed to mineralogical differences between samples. Samples outside the two shear zones show a gradual weakening trend, from $\mu \approx 0.6$ on the southwestern side to $\mu \approx 0.4$ on the northeastern side. This trend reflects a compositional change from more quartz-rich sandstone and siltstone on the southwestern side to more phyllosilicate-rich mudstones to the northeast (Supplementary Fig. 2). The SDZ marks the southwestern boundary of the damage zone, so that the samples with the highest residual strength reside outside the damage zone ($\mu \approx 0.50$–0.65).

The most significant strength observation is the abrupt decrease in μ within the two actively deforming shear zones. All residual strength measurements of the foliated gouge (at 10.4 mm slip) yield $\mu \leq 0.21$; the weakest sample has a strength of $\mu = 0.13$ (Fig. 1). A partly altered serpentinite porphyroclast from the SDZ has a strength of $\mu \approx 0.26$ and apparently survived by weaker matrix material flowing around it. The very low measured strengths are attributed to the abundance of the extremely weak mineral saponite ($\mu \approx 0.05$) (Fig. 2). Petrographic analysis indicates a saponite volume fraction of 60–65% in the foliated gouge matrix. Corrensite was also found in the SDZ and, based on composition, is likely to have a strength of $0.05 < \mu < 0.4$. Thus, corrensite along with increased quartz content (Supplementary Fig. 2) may be responsible for the marginally stronger frictional strength of the SDZ. Serpentine and minor amounts of other phyllosilicates (including chlorite, illite and micas) are present in the foliated gouge and, when added to the stronger quartz, feldspar and calcite constituents, result in a matrix strength that is consistent with estimates suggested by mixing law studies[11–13]. Rock fabric that localizes weak minerals can lower frictional strength relative to the strength of ground and mixed samples[14]. Thus, the SAFOD foliated gouge may be even weaker in its undisturbed state than the values reported here.

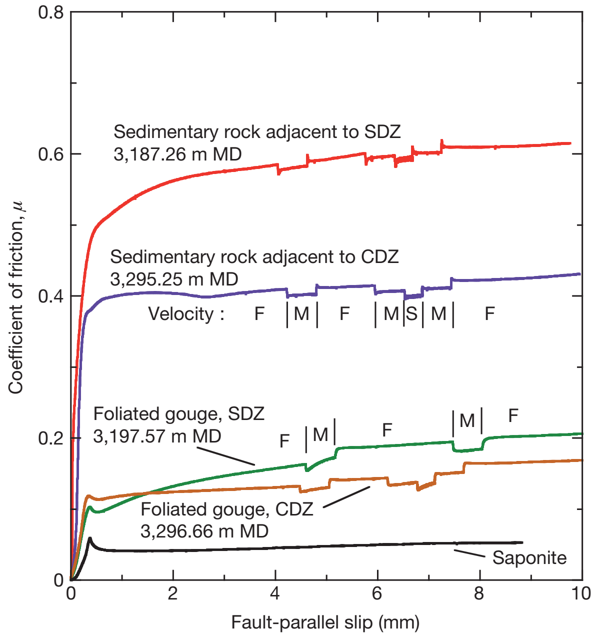


Figure 2 | Four representative deformation tests of core material from Hole G, with saponite for comparison. Periodic strength steps are due to decade changes in sliding rate (fault velocity in $\mu{\rm m\ s}^{-1}$: fast (F), 1.15; medium (M), 0.115; slow (S), 0.0115). Strength variations are attributed to compositional differences between samples as shown in XRD patterns. Bottom curve shows strength of monomineralic saponite gouge taken from vesicles in altered volcanic rocks from the Isle of Skye, Scotland (obtained from Mineralogical Research Co.). Permanent strengthening during some slow velocity steps is the result of time-dependent compaction. Foliated gouge is 3–4 times stronger than pure saponite owing to the presence of strong minerals like quartz. MD, measured depth.

Shear strength of fault gouge material typically varies with sliding rate. Rate dependence can be important in determining deformation mode (stable or unstable) and in extrapolating shear strength to SAF deformation rates. Steady-state rate sensitivity is defined[15] by the parameter $(a - b) = {\rm d}\mu_{\rm ss}/{\rm d}\ln V$, where $\mu_{\rm ss}$ is the steady-state friction coefficient at velocity V. Imposed velocity steps, as shown in Fig. 2, are used to determine $a - b$. Negative values promote unstable slip, whereas positive values are likely to result in stable creep. The serpentinite porphyroclast from the SDZ shows a range of both negative and positive values ($a - b = +0.0004 \pm 0.0014$) similar to past values reported for serpentinites[16,17]. All other core samples have positive rate sensitivity. Samples taken from outside the foliated gouge zones have values in the range $+0.001 < a - b < +0.007$. For CDZ, the combined measurements resulted in $a - b = +0.0018 \pm 0.0008$. For SDZ, values are twice as large: $a - b = +0.0037 \pm 0.0007$.

Average *in situ* strengths for CDZ and SDZ gouges (Fig. 1) are $\mu = 0.16$ and 0.19, respectively. These measurements are determined for a slip rate of $1.15\ \mu{\rm m\ s}^{-1}$ ($36{,}000\ {\rm mm\ yr}^{-1}$) and should be reduced to the appropriate *in situ* deformation rate ($\leq 34\ {\rm mm\ yr}^{-1}$) of the SAF. (Note that the slowest imposed deformation rate in the strength tests ($0.0115\ \mu{\rm m\ s}^{-1}$, Fig. 2) is only about 11 times the *in situ* rate.) Extrapolation of test strengths using the observed rate sensitivity for the foliated gouge indicates upper bounds for steady-state strength of CDZ and SDZ, respectively, of $\mu = 0.14$ and 0.16. This scaling assumes that the slip rate across the 1-mm-thick test gouge layer should be compared to the SAF deformation rate that is accommodated by the combined thickness of the SDZ and CDZ (~4.2 m). Depending on how strain is partitioned within the shear zones, the actual *in situ* shear strength supported by the deforming zones could be much less.

Although the SAF is one of the most well-studied fault systems in the world, fundamental questions about its strength and mechanical properties remain unanswered[18]. The SAF heat flow paradox was identified more than 40 years ago and is debated to this day[19–23]. Essentially, if the shear strength of the SAF were consistent with common laboratory-derived Byerlee rock friction ($\mu > 0.6$), frictional heating of the fault during earthquakes and stable fault creep should result in increased temperature and heat flow adjacent to the fault zone. In addition, the maximum horizontal stress near the fault should be oriented at ~30° to the fault trace. However, no evidence of a heat-flow anomaly along the creeping section of the SAF has been found[20,24], and borehole stress observations at SAFOD confirm that the maximum horizontal stress at this locality is at a high angle to the fault trace[4,5,25]. Although formation fluid pressure is apparently above hydrostatic in the sedimentary sequence northeast of the fault, there is no evidence from SAFOD drilling operations that pore pressure within the fault zone is elevated relative to the country rock[1]. The direct measurement, reported here, of low frictional strength ($\mu \approx 0.15$) of foliated gouge material taken at depth from the actively deforming shear zones is consistent with both the lack of an observed heat flow anomaly and the maximum compressive stress oriented at a high angle to the fault trace. Also, the positive dependence of strength on slip rate of the fault gouge material is consistent with deformation by creep rather than by earthquakes.

Saponite becomes unstable above about 150 °C (ref. 26) and is unlikely to be found deeper in the fault zone than 3.5–4 km (observed

©2011 Macmillan Publishers Limited. All rights reserved

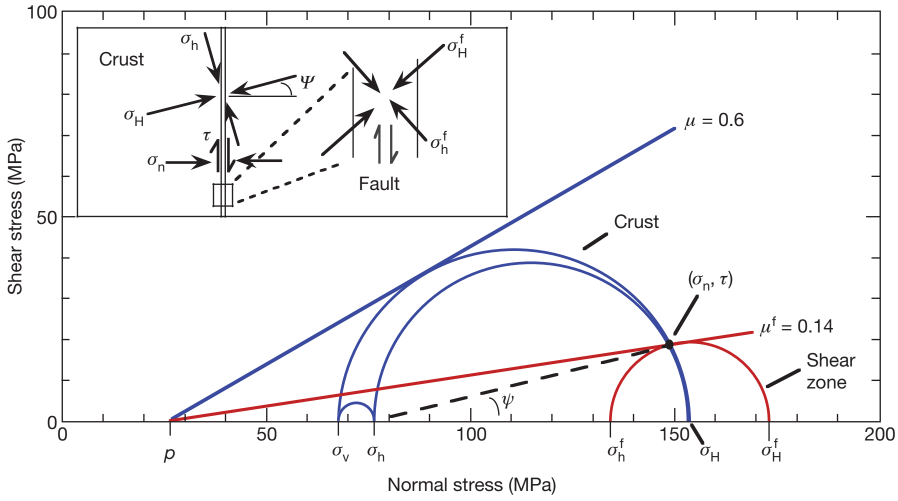


Figure 3 | Stress state for SAFOD drill site at 2.7 km depth. Model is based on borehole observations (assuming hydrostatic pore pressure) and foliated gouge strength (blue, host rock; red, weak shear zone). Main panel shows relationship of stress states within and outside the weak shear zone. Inset shows the corresponding spatial orientation of horizontal stresses in the model. While maximum shear stress in host rock is high, principal stresses rotate within the shear zone to accommodate the weaker material. In the fault, mean stress is high but shear stress is low. In model: $\sigma_H = 153$ MPa; $\sigma_h = 76$; $\sigma_v = 67.5$; $p = 27$. See text for definitions of symbols used.

temperature within the fault at ~2.7 km depth at SAFOD was 110–115 °C; C. Williams, personal communication). Stable creep and low strength of the deep SAF in the creeping section may reflect the presence of other low-strength minerals, elevated fluid pressure, or enhanced chemical weakening at greater depth than penetrated by SAFOD. Still, when considering the mechanics of the SAF specifically at 2.7 km, at SAFOD, mineralogy alone appears sufficient to explain fault strength.

Rice[27] analysed the stress state of a weak fault (due to elevated pore pressure, p) embedded in a stronger crust in a transpressional regime. Tembe *et al.*[28] extended the Rice analysis to include fault gouge of arbitrary strength. We use σ_H and σ_h for maximum and minimum horizontal principal stresses, respectively, and denote values within the fault by a superscript 'f'. Following Tembe *et al.*, a stress diagram representative of the SAFOD site at 2.7 km depth, with $\mu^f = 0.14$, $\mu = 0.60$ and hydrostatic p, is shown in Fig. 3. The model requires that σ_H (outside the weak shear zones) makes an angle of 77° to the strike of the SAF and is consistent with borehole observations at SAFOD showing high differential stresses in a transitional strike-slip to reverse-faulting stress regime, with σ_H maintaining a high angle to the SAF (70–80°) at depth[4,5]. Tractions on the fault, in this model, are $\tau = 17$ MPa and $\overline{\sigma_n} = 122$ MPa.

SAFOD phase 3 drilling has provided, for the first time, continuous core samples from the actively deforming SAF at a depth of 2.7 km. A self-consistent picture is emerging about the strength and deformation processes of this complex portion of the SAF that represents the transition from locked to creeping portions of the fault. Measurements of frictional strength of core material from within the SAF damage zone show two low-strength ($\mu \approx 0.15$) foliated gouge zones that are 1.6 and 2.6 m wide. These zones correspond to the actively creeping shear zones that were independently identified by casing deformation within the phase 2 hole. These shear zones are embedded in stronger material with $\mu \approx 0.35$–0.65. The extremely low strength of the foliated gouge in an otherwise strong crust is sufficient to explain the observed orientation of maximum compressive stress at a high angle relative to the strike of the fault (Fig. 3) without invoking high fluid pressure or other proposed fault-weakening mechanisms.

METHODS SUMMARY

We measured frictional strength of 25 samples obtained from the SAFOD phase 3 Hole G core, composed of material that could be carved from the core or removed as chips or rubble. Some portions of the rock mass bounding the shear zones had sufficient cohesion to be sampled as solid mini-cores or prisms and will be reported on later. Samples were ground to a powder (<150 µm diameter) and sheared in 1- or 2-mm-thick gouge layers between 25.4-mm-diameter driving blocks in a triaxial deformation apparatus[13] (Supplementary Fig. 1). Most samples were saturated with brine equivalent to *in situ* pore fluid (Y. Kharaka and J. Thordsen, personal communication) at 1 MPa constant pore pressure; a few tests were conducted with deionized water. Tests were performed at room temperature, constant effective normal stress (40, 120 and 200 MPa) and constant sliding rate (0.0115, 0.115 and 1.15 µm s^{-1}). Tests were carried out to 200 MPa to be applicable to *in situ* stress conditions (Fig. 3) and to allow for interpolation to other depths.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 19 October 2010; accepted 31 January 2011.
Published online 23 March 2011.

1. Zoback, M., Hickman, S. & Ellsworth, W. Scientific drilling into the San Andreas Fault zone. *Eos* **91,** 197–199 (2010).
2. Moore, D. E. & Rymer, M. J. Talc-bearing serpentinite, and the creeping section of the San Andreas fault. *Nature* **448,** 795–797 (2007).
3. Moore, D. E. & Rymer, M. J. Metasomatic origin of fault gouge comprising the two creeping strands at SAFOD. *Eos* (Fall suppl.), paper T41A-2105 (2010).
4. Boness, N. & Zoback, M. D. A multi-scale study of the mechanisms controlling shear velocity anisotropy in the San Andreas Fault Observatory at Depth. *Geophysics* **71,** F131–F146 (2006).
5. Hickman, S. & Zoback, M. D. Stress measurements in the SAFOD pilot hole: implications for the frictional strength of the San Andreas fault. *Geophys. Res. Lett.* **31,** L15S12, doi:10.1029/2004GL020043 (2004).
6. Zoback, M. D., Hickman, S. & Ellsworth, W. in *Treatise on Geophysics* Vol. 4 (ed. Schubert, G.) 649–674 (Elsevier, 2007).
7. Tobin, H., Ito, H., Behrmann, J., Hickman, S. H. & Kimura, G. in *Report from IODP/ICDP Workshop on Fault Zone Drilling, Scientific Drilling* (eds Ito, H. *et al.*) 5–16 (Special Issue No. 1, Integrated Ocean Drilling Program, Hokkaido, 2007).
8. Solum, J. G. *et al.* Mineralogical characterization of protolith and fault rocks from the SAFOD main hole. *Geophys. Res. Lett.* **33,** L21314, doi:10.1029/2006GL027285 (2006).
9. Tembe, S. *et al.* Frictional strength of cuttings and core from SAFOD drillhole phases 1 and 2. *Geophys. Res. Lett.* **33,** L23307, doi:10.1029/2006GL027626 (2006).
10. Holdsworth, R. E. *et al.* Fault rocks from the SAFOD core samples: implications for weakening at shallow depths along the San Andreas Fault, California. *J. Struct. Geol.* **33,** 132–134 (2011).
11. Moore, D. E. & Lockner, D. A. Frictional strengths of talc-serpentinite and talc-quartz mixtures. *J. Geophys. Res.* **116,** B01403, doi:10.1029/2010JB007881 (2011).
12. Crawford, B. R., Faulkner, D. R. & Rutter, E. H. Strength, porosity, and permeability development during hydrostatic and shear loading of synthetic quartz-clay fault gouge. *J. Geophys. Res.* **113,** B03207, doi:10.1029/2006JB004634 (2008).
13. Tembe, S., Lockner, D. A. & Wong, T.-f. Effect of clay content and mineralogy on frictional sliding behavior of simulated gouges: binary and ternary mixtures of quartz, illite and montmorillonite. *J. Geophys. Res.* **115,** B03416, doi:10.1029/2009JB006383 (2010).
14. Collettini, C., Niemeijer, A., Viti, C. & Marone, C. Fault zone fabric and fault weakness. *Nature* **462,** 907–910 (2009).
15. Dieterich, J. H. Modeling of rock friction 1. Experimental results and constitutive equations. *J. Geophys. Res.* **84,** 2161–2168 (1979).

©2011 Macmillan Publishers Limited. All rights reserved

16. Moore, D. E., Lockner, D. A., Ma, S., Summers, R. & Byerlee, J. D. Strengths of serpentinite gouges at elevated temperatures. *J. Geophys. Res.* **102,** 14787–14801 (1997).
17. Moore, D. E., Lockner, D. A., Tanaka, H. & Iwata, K. The coefficient of friction of chrysotile gouge at seismogenic depths. *Int. Geol. Rev.* **46,** 385–398 (2004).
18. Hickman, S. H. Stress in the lithosphere and the strength of active faults. *Rev. Geophys.* **29,** 759–775 (1991).
19. Brune, J. N., Henyey, T. L. & Roy, R. F. Heat flow, stress, and rate of slip along the San Andreas fault, California. *J. Geophys. Res.* **74,** 3821–3827 (1969).
20. Lachenbruch, A. H. & Sass, J. H. Heat flow and energetics of the San Andreas fault zone. *J. Geophys. Res.* **85,** 6185–6222 (1980).
21. Scholz, C. H. Evidence for a strong San Andreas fault. *Geology* **28,** 163–166 (2000).
22. Zoback, M. D. Strength of the San Andreas. *Nature* **405,** 31–32 (2000).
23. Saffer, D. M., Bekins, B. A. & Hickman, S. Topographically driven groundwater flow and the San Andreas heat flow paradox revisited. *J. Geophys. Res.* **108,** 2274, doi:10.1029/2002JB001849 (2003).
24. Lachenbruch, A. H. & Sass, J. H. Heat flow from Cajon Pass, fault strength, and tectonic implications. *J. Geophys. Res.* **97,** 4995–5015 (1992).
25. Zoback, M. D. *et al.* New evidence on the state of stress of the San Andreas fault system. *Science* **238,** 1105–1111 (1987).
26. Inoue, A. & Utada, M. Smectite-to-chlorite transformation in thermally metamorphosed volcanoclastic rocks in the Kamikita area, northern Honshu, Japan. *Am. Mineral.* **76,** 628–640 (1991).
27. Rice, J. R. in *Fault Mechanics and Transport Properties of Rocks* (eds Evans, B. & Wong, T.-f.) 475–503 (Academic, 1992).
28. Tembe, S., Lockner, D. & Wong, T.-f. Constraints on the stress state of the San Andreas fault with analysis based on core and cuttings from San Andreas Observatory at Depth (SAFOD) drilling phases 1 and 2. *J. Geophys. Res.* **114,** B11401, doi:10.1029/2008JB005883 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. Weymer, J. Firth and P. Nelson for their assistance and patience in providing the core material. Comments from W. Ellsworth, D. Faulkner and R. Simpson significantly improved the manuscript.

Author Contributions C.M. and D.A.L. conducted deformation experiments, D.M. provided petrographic analysis and XRD measurements, and S.H. provided borehole logging results and related stress analysis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.A.L. (dlockner@usgs.gov).

©2011 Macmillan Publishers Limited. All rights reserved

METHODS

Sample preparation. We measured frictional strength of 25 samples that were obtained from the SAFOD phase 3 Hole G core. Hole G was cored in measured depth intervals 3,186.7–3,199.5 m and 3,294.9–3,312.7 m to sample localized shear zones within the SAF damage zone that correspond to the two intervals, referred to as SDZ and CDZ, where slow deformation was observed in the Phase 2 casing[1]. As indicated in Fig. 1, the cored intervals in Hole G were within or adjacent to the SAF damage zone as determined by logging data following phase 2 drilling. While selected samples were obtained to provide whole, undisturbed wafers for intact strength tests, samples tested here were either carved from the core or collected as rubble, chips or loose powder. All samples reported here were prepared by repeated gentle grinding with mortar and pestle until all material passed through a 100 mesh sieve (0.15 mm diameter). Resulting powder was then wetted to make a paste that was formed into a 1-mm-thick test layer (2-mm layers were used in the 200 MPa tests). The first 14 tests were performed with deionized water. All remaining tests used a prepared brine solution that duplicates the major cations and their relative concentrations found in the formation fluid retrieved from the SAFOD drill hole on the northeastern side of the SAF (Y. Kharaka and J. Thordsen, personal communication). Test fluid constituents, expressed in units of grams per litre, are: Cl^-, 13.32; Na^+, 5.34; Ca^{2+}, 2.77; and K^+, 0.22. Comparison tests showed only a slight difference between frictional strength for samples sheared with deionized water and samples sheared with the brine solution. Before mechanical testing, X-ray diffraction patterns were obtained to determine mineral composition and relative abundance.

Testing details. Tests were performed in a standard triaxial apparatus at room temperature and effective normal stresses of 40, 120 and 200 MPa. A constant pore pressure of 1 MPa was applied in all tests. Samples were 25.4-mm-diameter right-cylinders that contained a sawcut inclined 30° to the sample axis (Supplementary Fig. 1). The sawcut forcing blocks were sandstone–sandstone, sandstone–granite, or granite–granite pairs. Surfaces of forcing blocks were roughened with 100 grit abrasive, to assure good frictional contact with the applied gouge layer. See, for example, refs 9 and 13 for details. Berea sandstone forcing blocks have high permeability but also have 20% porosity that decreases with applied load. The standard test geometry used Berea for the top forcing block to assure rapid hydraulic communication of the fault with the external pore fluid system. Pore fluid flow in and out of the lower driving block was through the fault gouge. To minimize pore pressure transients due to stress changes, low-porosity granite was used for the lower driving block in most tests, and particularly in experiments with low-permeability clay-rich gouge.

A greased Teflon shim, placed between the piston and the sample assembly, allowed easy lateral slip of the lower driving block during shearing (Supplementary Fig. 1). Samples were jacketed in 3.2-mm-thick latex tubing for isolation from confining fluid. Separate calibration tests showed that the latex tubing provided an equivalent shear resistance of 0.043 MPa mm^{-1} due to stretching during deformation experiments. This shear strength correction has been applied to all test results. Shear and normal stresses have also been corrected for the reduction in contact area as the two sample halves slide past each other[13]. Axial load was measured with an internal load cell. Axial shortening, confining pressure and pore pressure were all measured at 1-s intervals. Shear and normal stress resolved on the fault surface were computed in real time from the axial stress, confining pressure and axial shortening. As sample strength varied, confining pressure was adjusted every second to maintain constant normal stress. Axial stress, confining pressure and pore pressure have accuracies of at least 0.03 MPa. Samples were sheared to 9 mm axial shortening (~10.4 mm parallel to the sawcut) at axial shortening rates of 0.01, 0.1 and 1.0 $\mu\mathrm{m\,s}^{-1}$ to determine the dependence of shear strength on sliding rate and thereby the tendency for stable creep or unstable slip. Slip and slip rate on the inclined fault surfaces were 15% higher than the corresponding axial values. Steady-state changes in strength were estimated for individual velocity steps by measuring the residual strength change after de-trending the friction–displacement curves for long-term strain hardening. This procedure was carried out manually.

Measurement accuracy. Sample strength is reported as coefficient of friction $\mu = \tau/\overline{\sigma_{\mathrm{n}}}$. Within a single experiment, changes in μ have a precision of ±0.001. Reproducibility of μ between experiments, including variations due to sample preparation, is approximately ±0.005. Accuracy of μ, after corrections for true contact area and jacket strength, is approximately ±0.01. Initial gouge layer thickness is 1.0 mm for 40 and 120 MPa tests. Gouge layer thickness is 2.0 mm for 200 MPa tests to offset thinning at the higher normal stress. Compaction is not measured during experiments, but layer thickness following experiments is reduced by 5–30%, depending on normal stress, gouge clay content and driving block type. As shear will localize within the gouge layer to different degrees and at different times, depending on composition and normal stress, estimates of true shear strain are problematic. The deformation quantity that is most accurately determined in these experiments is total fault-parallel slip. This can be converted to a nominal shear strain by dividing by the initial 1 or 2 mm gouge thickness.

©2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09904

Biodiversity improves water quality through niche partitioning

Bradley J. Cardinale[1]

Excessive nutrient loading of water bodies is a leading cause of water pollution worldwide[1,2], and controlling nutrient levels in watersheds is a primary objective of most environmental policy[3]. Over the past two decades, much research has shown that ecosystems with more species are more efficient at removing nutrients from soil and water than are ecosystems with fewer species[4–7]. This has led some to suggest that conservation of biodiversity might be a useful tool for managing nutrient uptake and storage[7–10], but this suggestion has been controversial, in part because the specific biological mechanisms by which species diversity influences nutrient uptake have not been identified[10–12]. Here I use a model system of stream biofilms to show that niche partitioning among species of algae can increase the uptake and storage of nitrate, a nutrient pollutant of global concern. I manipulated the number of species of algae growing in the biofilms of 150 stream mesocosms that had been set up to mimic the variety of flow habitats and disturbance regimes that are typical of natural streams. Nitrogen uptake rates, as measured by using ^{15}N-labelled nitrate, increased linearly with species richness and were driven by niche differences among species. As different forms of algae came to dominate each unique habitat in a stream, the more diverse communities achieved a higher biomass and greater ^{15}N uptake. When these niche opportunities were experimentally removed by making all of the habitats in a stream uniform, diversity did not influence nitrogen uptake, and biofilms collapsed to a single dominant species. These results provide direct evidence that communities with more species take greater advantage of the niche opportunities in an environment, and this allows diverse systems to capture a greater proportion of biologically available resources such as nitrogen. One implication is that biodiversity may help to buffer natural ecosystems against the ecological impacts of nutrient pollution.

Over the past century, humans have more than doubled the rate of nitrogen input into terrestrial ecosystems, mostly through fossil fuel combustion and increased use of agricultural fertilizers[1,2]. Excess nitrogen flows into streams and rivers, where it contributes to eutrophication, one of the leading causes of degraded water quality worldwide[13,14]. If it is not removed by biotic uptake or denitrification, nitrogen is ultimately exported from rivers to estuaries and coastal oceans[15,16], where it can promote blooms of harmful microorganisms[17] and can generate an excessive biochemical oxygen demand, which has resulted in 245,000 km^2 of 'dead zones' in more than 400 coastal habitats around the world[18]. Given the global ecological and economic impact of these environmental problems[19], a primary objective of environmental policy and management is to control the input of nitrogen into watersheds and to maximize its removal[3].

A growing body of recent research has suggested that conservation of biodiversity might be a useful management tool for reducing the concentrations of nitrogen in soil and water[4–10]. It is often argued that communities with more species take greater advantage of the niche opportunities that are available in an environment than do those with fewer species, and this allows diverse ecosystems to capture a greater proportion of biologically active resources such as nitrogen[20,21]. But the role of niche differences among species in regulating nutrient uptake has rarely been measured directly[6,12,22]. Instead, effects of niche differences on nutrient uptake either have been assumed to occur based on theoretical arguments[20,23] or have been inferred from post hoc statistical analyses of experiments that have been unable to differentiate among specific biological mechanisms[24]. Because of the lack of direct evidence confirming a biological mechanism, there has been considerable debate about why diverse ecosystems tend to be more efficient at sequestering dissolved nutrients[10–12].

Here I present the results of a novel experiment in which I directly manipulated algae and their niche opportunities in a large set of stream mesocosms to examine the mechanistic links between the diversity of algae, niche differences among species and the rate of nitrate removal from stream water. I isolated cultures of eight of the most widespread species of diatoms and green algae that inhabit streams in North America: five species of Bacillariophyceae (the diatoms *Achnanthidium minutissimum*, *Melosira varians*, *Navicula cryptocephala*, *Nitzschia palea* and *Synedra ulna*), and three chlorophyceaen green algae (*Scenedesmus quadricauda*, *Stigeoclonium* sp. and *Spirogyra* sp.) (Supplementary Tables 1 and 2). These species were then used to colonize recirculating stream channels with equal initial cell densities of one, two, four, six or eight species (see Methods). Streams were set up to mimic two forms of environmental heterogeneity that are typical of natural fluvial ecosystems and that are thought to provide niche opportunities allowing species to coexist. Flow heterogeneity was created by arranging algal growth surfaces so that velocities near the stream bed varied from <2 to 55 $cm\ s^{-1}$. The abundance, distribution and diversity of stream algae are known to be influenced by a stream's flow regime[25], and species have evolved varying morphological adaptations to cope with near-bed shear stress[26,27]. In addition to spatial heterogeneity in flow, temporal heterogeneity was created by dividing the growth surfaces in a stream into 18 distinct habitat patches ($2.5 \times 2.5\ cm^2$), each of which had a probability of 0.25 of being disturbed (by removal with a brush) in each week of the 6-week experiment. This treatment produced a successional mosaic of patches that varied from 5 to 50 days old by the end of the experiment. Different species often dominate habitats at different stages of succession owing to trade-offs between the ability of species to colonize the available space versus their ability to compete with other species. These trade-offs can enhance coexistence in streams that experience periodic disturbances[28].

In heterogeneous streams, nitrogen uptake rates were a linear function of species diversity across the range of richness used in this study, increasing by 0.06 $\mu g\ NO_3^-\ cm^{-2}\ h^{-1}$ for each species of alga in the biofilm (Fig. 1a and Table 1). The most diverse algal polyculture (eight species) removed nitrate 4.5-fold faster than the average rate of a species grown alone (in monoculture). Rates of uptake in streams colonized with the eight-species polyculture were also significantly faster than the rates achieved by the single most efficient species used in the study (95% confidence interval for polycultures, 1.45–4.01-fold faster than *Stigeoclonium*) (Fig. 1a, dashed line).

Species diversity did not alter the biomass-specific rates of nutrient uptake. There was no difference in the NO_3^- uptake rate per unit of

[1] University of Michigan, School of Natural Resources & Environment, Ann Arbor, Michigan 48109-1041, USA.

©2011 Macmillan Publishers Limited. All rights reserved

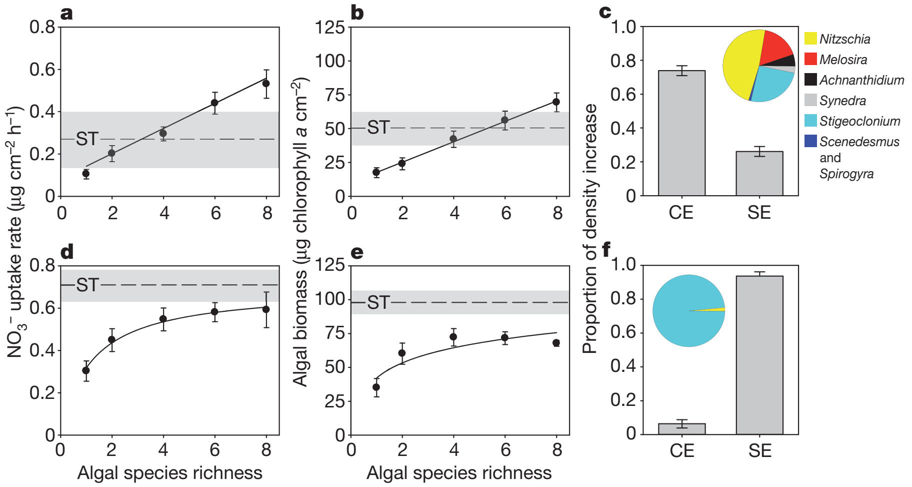


Figure 1 | Algal diversity effects on NO$_3^-$, algal biomass and final population sizes. a–c, Heterogeneous streams, with flow varying spatially and habitats varying in successional age. **d–f**, Homogeneous streams, in which niche opportunities had been removed. Data are presented as mean ± s.e.m. of 24 replicates for monocultures, 15 replicates for 2–6 species polycultures and 6 replicates for 8-species polycultures. Best fitting functions (Table 1) are plotted as solid lines. The horizontal line and the grey shaded area show mean ± s.e.m. for *Stigeoclonium*, which achieved the highest values of all of the monocultures. **c**, **f**, The proportion of increased polyculture cell densities driven by niche complementarity (CE) or selection effects (SE; that is, the influence of dominant species).

chlorophyll across the levels of richness ($P = 0.38$, linear mixed effects model, see Methods). However, a significant proportion of the variation in NO$_3^-$ uptake could be explained by differences in the total algal biomass among streams (μg N cm^{-2} h^{-1} = 0.08 + (0.01 × μg chlorophyll *a* cm^{-2}), $P < 0.01$, $R^2 = 0.56$). Algal biomass was a linear function of diversity across the range of richness used in the study, increasing by 7.67 μg chlorophyll *a* cm^{-2} for each additional species (Table 1 and Fig. 1b). Increased algal biomass, and consequently the higher rate of NO$_3^-$ uptake, led to a strong relationship between algal species richness and the total amount of nitrogen stored in the biofilm at the end of the experiment (μg [^{15}N + ^{14}N] cm^{-2} = 45 + (15 × richness), $P < 0.01$, $R^2 = 0.45$).

Two lines of evidence suggest that the effects of algal diversity on NO$_3^-$ uptake and storage were driven by niche differences among species. First, if the frequency of disturbance and the spatial variation in flow represent axes of a species niche, then ecological theory predicts that different species should dominate different areas of this two-dimensional niche space[20,23]. Detailed examination of the algal population sizes showed that different morphological forms of algae dominated unique, complementary habitats in the streams (Fig. 2). High-velocity habitats were dominated by single-celled diatoms that grow prostrate to a surface in a way that is resistant to displacement by shear (for example, *Achnanthidium* and *Synedra*). By contrast, low-velocity habitats were dominated by large, filamentous algae that are susceptible to shear (for example, *Stigeoclonium* and *Melosira*). Fast-growing diatoms (such as *Achnanthidium* and *Nitzschia*) dominated habitats that had experienced recent disturbance. These early successional species were replaced by larger colonial, filamentous or slow-growing species in habitats that were less frequently disturbed (for example, *Spirogyra*, *Stigeoclonium* and *Synedra*). Differential habitat use by species led to a phenomenon called ‘over-yielding’ in algal polycultures, with five species achieving higher cell densities than would be expected on the basis of their initial proportional density (see Methods). As a result, more than 80% of increased cell densities in polycultures were driven by niche complementarity (Fig. 1c), which is enhanced when species use habitats or resources in ways that are either unique or synergistic[24].

The second line of evidence that suggests niche differences are responsible for increased NO$_3^-$ uptake stems from the findings when niche opportunities were experimentally removed. Ecological theory predicts that when niche opportunities are eliminated from a system, the effects of biodiversity should disappear, and ecological processes should become dominated by a single, competitively superior species[20]. This is exactly what was observed. When the same algal species were grown in streams that had been forced to be spatially homogeneous (with near-bed flow velocity set to a uniform 22 cm s^{-1}, the median velocity of heterogeneous streams) and in which patches were never disturbed over the course of the experiment (all patches ~50 days old, with no variation in successional age), the previously observed effect of algal species richness on NO$_3^-$ uptake rates disappeared. In these

Table 1 | Effects of algal diversity on rates of NO$_3^-$ uptake and algal biomass

Dependent variable	Heterogeneous streams, with niche opportunities*				Homogeneous streams, with niche opportunities removed†			
	a	*b*	AIC	W_i	*a*	*b*	AIC	W_i
NO$_3^-$ uptake rate (μg cm^{-2} h^{-1})								
Linear, $y = a + b \times S$	**0.05**	**0.06**	**−59.99**	**0.58**	0.34	0.05	8.73	0.00
Log, $y = a + b \times \log(S)$	0.09	0.19	−59.32	0.42	0.35	0.17	2.50	0.04
Hyperbolic, $y = a \times S/(b + S)$	0.06	−3.34	32.54	0.00	**0.77**	**1.17**	**−4.13**	**0.96**
Algal biomass (μg chlorophyll *a* cm^{-2})								
Linear, $y = a + b \times S$	**9.77**	**7.67**	**663.39**	**0.62**	42.79	6.11	716.44	0.02
Log, $y = a + b \times \log(S)$	14.30	22.58	664.41	0.37	**43.23**	**21.41**	**708.53**	**0.98**
Hyperbolic, $y = a \times S/(b + S)$	115.84	6.57	674.23	0.00	97.23	1.23	720.77	0.00

Measurements were taken in heterogeneous streams, where spatial variation in flow and periodic disturbances generated unique ecological niche opportunities, and homogenous streams, where niche opportunities had been intentionally removed by making all habitats physically uniform. Data were fit to linear, decelerating (log) and saturating (Michaelis–Menten hyperbolic) functions to characterize the general form of the relationship between algal richness and each response variable. The Akaike information criterion (AIC) was used to estimate Akaike weights, W_i, which give the relative likelihood of each model given the data. Highest relative likelihoods are in bold. *S*, species richness of algae.

* See also Fig. 1a–c.

† See also Fig. 1d–f.

©2011 Macmillan Publishers Limited. All rights reserved

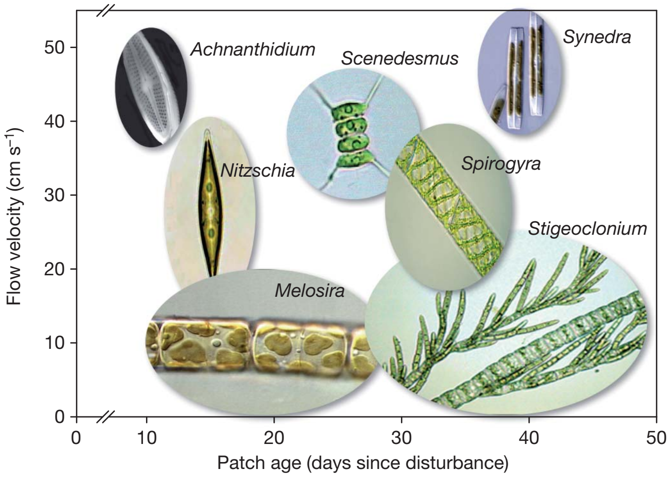


Figure 2 | Niche partitioning by algae. The ovals show the mean (centre of image) ± 95% confidence interval (boundary of image) of cell densities along two axes of a species niche (successional age of habitat and near-bed velocity). Filamentous algae that are susceptible to shear (*Melosira* and *Stigeoclonium*) were abundant in low-velocity habitats. Single-celled diatoms that grow prostrate to a surface (*Achnanthidium* and *Synedra*) achieved the highest densities in high-velocity habitats. Early successional habitats were dominated by small diatoms with fast rates of growth (*Achnanthidium* and *Nitzschia*), whereas late successional habitats were dominated by slow-growing cells, colonies or filaments (*Stigeoclonium*, *Spirogyra* and *Synedra*). *Navicula* is not plotted, because it failed to establish itself in polyculture despite growing in monoculture.

homogeneous streams, NO_3^- uptake rates and algal biomass were both positive but quickly decelerating functions of algal richness (Fig. 1d, e and Table 1). The most diverse polycultures took up NO_3^- at a rate that was twofold faster than the average monoculture, but these rates were significantly slower than that of the single most efficient species grown in monoculture (95% confidence interval for polycultures, -0.18 to $-0.21 \times$ values for *Stigeoclonium*, $P = 0.04$, one-sided *t*-test). In these streams, more than 90% of the increased cell density in polycultures was driven by species-specific selection effects, which occur when initially diverse systems become dominated by a single, competitively superior taxon (in this case the filamentous green alga *Stigeoclonium*) (Fig. 1f). Such results suggest that in a homogeneous environment, the loss of niche opportunities led to reduced diversity and caused NO_3^- uptake to be controlled more by a single species than by algal species richness.

The results of this study build on previous research efforts by providing direct experimental evidence that ecological niche differences among species allow diverse communities to increase their uptake and tissue storage of nitrate. The specific linear responses documented in this study might not be expected to extrapolate to conditions in the field, where experiments performed with more species often show nonlinear responses[22]. Even so, this study is important because it confirms one of the fundamental mechanisms that has long been presumed to underlie the effects of diversity on ecological processes in ecosystems that range from the simple to the complex. That mechanism—niche partitioning—arose in this system because different species were best adapted for different habitats in a stream. These adaptations were expressed only when environmental conditions were dynamic in space and/or time and when heterogeneity provided ecological opportunities for species to coexist. Small, adnate forms of algae, which are known to be resistant to shear, dominated high-flow environments, whereas large, filamentous species that are prone to shear dominated low-flow environments. Fast-growing species dominated disturbed habitats, whereas slow-growing, competitively superior species dominated late successional habitats. Differences in ecological form allowed species to occupy unique and complementary habitat types in a stream. In turn, streams with diverse communities of microalgae had a greater capacity to remove NO_3^- than did streams with fewer numbers of species.

NO_3^- is one of the most abundant pollutants worldwide[13,14]; therefore, one implication of this study is that biodiversity could have a role in sequestering pollutants from natural environments. However, an important caveat to this conclusion is that nutrient pollution itself is known to reduce biodiversity, both through loss of species and through increased dominance of certain types of primary producer (for example, *Cladophora*). As a result, there is the potential for dual causality, whereby high nutrient loading of ecosystems reduces biodiversity, while the existing diversity reduces nutrient concentrations. How this feedback balances out to control diversity and nutrient levels in rivers that are increasingly being homogenized by human activities (for example, damming and altered flow regimes[29]) is an area of active research[30]. Nevertheless, the results of this study suggest two points. First, in those ecosystems where high nutrient loading reduces diversity, attempts to restore nutrient cycling to pre-disturbance conditions will probably be hampered by the irreversible loss of species. Second, in those ecosystems where it is possible to maintain species diversity despite nutrient loading, biodiversity may help to buffer ecosystems against the ecological impacts of nutrient pollution. Buffering against nutrient pollution will require not only the conservation of biodiversity, but also conservation of the forms of environmental heterogeneity that create niche opportunities and allow species to coexist.

METHODS SUMMARY

The species used for this study included eight forms of diatom and green alga that are among the most widespread and abundant species in North American streams (Supplementary Tables 1 and 2). The experiment was performed in the stream flume facility at the University of California Santa Barbara. This facility contains 120 recirculating laboratory 'flumes' (Supplementary Fig. 1), which are widely used for laboratory studies of lotic algae because they allow high replication rates, long-term population dynamics and large population sizes (Supplementary Information).

Algal diversity (one, two, four, six or eight species) was manipulated as a substitutive design (constant initial cell density of 10,000) in flumes that had two types of growth environment: heterogeneous and homogeneous. Heterogeneous flumes were constructed such that habitats within a stream had a wide range of near-bed velocities (from $<2\,\mathrm{cm\,s^{-1}}$ to $55\,\mathrm{cm\,s^{-1}}$), as well as disturbance regimes that allowed habitats to differ in successional age (from 5 to 50 days old). Homogeneous streams were constructed so that near-bed flow velocity was uniform ($22\,\mathrm{cm\,s^{-1}}$, the median value for heterogeneous streams), and habitat patches were never disturbed. Thus, heterogeneous flumes had two forms of environmental variation that provided ample opportunity for species to express niche differences. By contrast, homogeneous flumes had no obvious spatial or temporal variation that might allow the expression of niche differences.

After allowing biofilms to reach a steady-state biomass, ^{15}N-enriched $NaNO_3$ (60% ^{15}N) was added to flumes in concentrations sufficient to increase $^{15}N/^{14}N$ ratios in the dissolved pool by 1.88-fold, but changing total NO_3^- by just 1.004-fold. After incubation, biofilms were destructively sampled to measure the amount of ^{15}N that had been sequestered by the biofilm and to estimate the final sizes of the algal populations.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 8 November 2010; accepted 7 February 2011.

1. Vitousek, P. M. *et al.* Human alteration of the global nitrogen cycle: sources and consequences. *Ecol. Appl.* **7,** 737–750 (1997).
2. Canfield, D. E., Glazer, A. N. & Falkowski, P. G. The evolution and future of Earth's nitrogen cycle. *Science* **330,** 192–196 (2010).
3. Smith, V. H. & Schindler, D. W. Eutrophication science: where do we go from here? *Trends Ecol. Evol.* **24,** 201–207 (2009).
4. Spehn, E. M. *et al.* Ecosystem effects of biodiversity manipulations in European grasslands. *Ecol. Monogr.* **75,** 37–63 (2005).
5. Cardinale, B. J. *et al.* Effects of biodiversity on the functioning of trophic groups and ecosystems. *Nature* **443,** 989–992 (2006).
6. Bracken, M. E. S. & Stachowicz, J. J. Seaweed diversity enhances nitrogen uptake via complementary use of nitrate and ammonium. *Ecology* **87,** 2397–2403 (2006).

©2011 Macmillan Publishers Limited. All rights reserved

7. Tilman, D., Wedin, D. & Knops, J. Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature* **379,** 718–720 (1996).
8. Scherer-Lorenzen, M., Palmborg, C., Prinz, A. & Schulze, E. D. The role of plant diversity and composition for nitrate leaching in grasslands. *Ecology* **84,** 1539–1552 (2003).
9. Reich, P. B. *et al.* Plant diversity enhances ecosystem responses to elevated CO_2 and nitrogen deposition. *Nature* **410,** 809–812 (2001).
10. Hooper, D. U. *et al.* Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecol. Monogr.* **75,** 3–35 (2005).
11. Huston, M. A. Hidden treatments in ecological experiments: re-evaluating the ecosystem function of biodiversity. *Oecologia* **110,** 449–460 (1997).
12. Loreau, M. *et al.* Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science* **294,** 804–808 (2001).
13. Carpenter, S. R. *et al.* Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecol. Appl.* **8,** 559–568 (1998).
14. Dodds, W. K. Eutrophication and trophic state in rivers and streams. *Limnol. Oceanogr.* **51,** 671–680 (2006).
15. Mulholland, P. J. *et al.* Stream denitrification across biomes and its response to anthropogenic nitrate loading. *Nature* **452,** 202–205 (2008).
16. Alexander, R. B., Smith, R. A. & Schwarz, G. E. Effect of stream channel size on the delivery of nitrogen to the Gulf of Mexico. *Nature* **403,** 758–761 (2000).
17. Anderson, D. M., Glibert, P. M. & Burkholder, J. M. Harmful algal blooms and eutrophication: nutrient sources, composition, and consequences. *Estuaries* **25,** 704–726 (2002).
18. Diaz, R. J. & Rosenberg, R. Spreading dead zones and consequences for marine ecosystems. *Science* **321,** 926–929 (2008).
19. Dodds, W. K. *et al.* Eutrophication of US freshwaters: analysis of potential economic damages. *Environ. Sci. Technol.* **43,** 12–19 (2009).
20. Tilman, D., Lehman, D. & Thompson, K. Plant diversity and ecosystem productivity: theoretical considerations. *Proc. Natl Acad. Sci. USA* **94,** 1857–1861 (1997).
21. Chapin, F. S. *et al.* Biotic control over the functioning of ecosystems. *Science* **277,** 500–504 (1997).
22. Cardinale, B. J. *et al.* The functional role of producer diversity in ecosystems. *Am. J. Bot.* **98,** 572–592 (2011).
23. Chesson, P. Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* **31,** 343–366 (2000).
24. Loreau, M. & Hector, A. Partitioning selection and complementarity in biodiversity experiments. *Nature* **412,** 72–76 (2001).
25. Peterson, C. G. & Stevenson, R. J. Resistance and resilience of lotic algal communities: importance of disturbance timing and current. *Ecology* **73,** 1445–1461 (1992).
26. Biggs, B. J. F. & Thomsen, H. A. Disturbance of stream periphyton by perturbations in shear stress: time to structural failure and differences in community resistance. *J. Phycol.* **31,** 233–241 (1995).
27. Steinman, A. D. & McIntire, C. D. Effects of current velocity and light energy on the structure of periphyton assemblages in laboratory streams. *J. Phycol.* **22,** 352–361 (1986).
28. Pringle, C. M. Patch dynamics in lotic systems: the stream as a mosaic. *J. N. Am. Benthol. Soc.* **7,** 503–524 (1988).
29. Poff, N. L., Olden, J. D., Merritt, D. M. & Pepin, D. M. Homogenization of regional river dynamics by dams and global biodiversity implications. *Proc. Natl Acad. Sci. USA* **104,** 5732–5737 (2007).
30. Cardinale, B. J., Bennett, D. M., Nelson, C. E. & Gross, K. Does productivity drive diversity or vice versa? A test of the multivariate productivity–diversity hypothesis in streams. *Ecology* **90,** 1227–1241 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Byrnes, W. Dodds, J. Levine and A. Steinman for comments that improved the manuscript. D. Bennett, K. Matulich, L. Power and J. Weis helped to set up and run the experiment. M. Potapova provided images of the algae in Fig. 2. This work was funded by grants from the US National Science Foundation (DEB 0614428 and DEB 1046121).

Author Contributions B.J.C. designed the study, collected data with the assistance of those mentioned in the Acknowledgements, analysed the data and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.J.C. (bradcard@umich.edu).

©2011 Macmillan Publishers Limited. All rights reserved

METHODS

Species pool. The algae used in the experiment included five species of Bacillariophyceae (the diatoms *Achnanthidium minutissimum*, *Melosira varians*, *Navicula cryptocephala*, *Nitzschia palea* and *Synedra ulna*) and three Chlorophyceaen green algae (*Scenedesmus quadricauda*, *Stigeoclonium* sp. and *Spirogyra* sp.). Although these taxa do not necessarily represent a random subset of all stream algae (a limited proportion of algal species exist in culture), they do rank among the most common and abundant types of primary producers in rivers throughout North America (Supplementary Table 1). In addition, representatives of these genera co-occur across a large proportion of rivers (Supplementary Table 2). As such, the study is representative of algal species that co-occur in many streams.

These species have a variety of morphologies, which are thought to influence the type of habitat the species occupy (Supplementary Fig. 1 and Supplementary Table 1). For example, *Achnanthidium* and *Synedra* are single-celled diatoms that attach prostrate to a growth surface, a function that is thought to be an adaptation to living in high shear environments[26,31,32]. By contrast, *Melosira*, *Stigeoclonium* and *Spirogyra* are large, filamentous algae that are known to be susceptible to drag and displacement by high shear stress[26,31]. Some of the species have fast rates of cell division and a high potential for colonization (for example, *Achnanthidium* and *Nitzschia*); as such, they often dominate early successional habitats after a disturbance[25,27,33,34]. Others are slow growing and tend to dominate late successional communities (for example, the diatom *Synedra*, the colonial alga *Scenedesmus* or the filamentous alga *Stigeoclonium*)[25,27,33].

Stream mesocosms. Algal diversity was manipulated in recirculating laboratory flumes ($0.5 \times 0.1 \times 0.1\ m^3$) in which discharge was controlled by 7-cm diameter propellers driven by a d.c. motor attached to a Hy3020E 3-A voltage controller (TekPower) (Supplementary Fig. 1). Biofilms in each flume were grown on a 200-cm^2 polyvinyl chloride (PVC) growth surface that had been roughened by sanding to facilitate colonization and growth. Growth surfaces were placed under a T5 Aquarium lighting fixture (Coralife) containing two 9-W, 10-K daylight-spectrum fluorescent lamps (Supplementary Fig. 1).

At the start of the experiment, each flume was sterilized with bleach and then filled with 13 l of 10% sterile Chu culture medium. Chu medium is widely used for growing freshwater algae and contains a suite of macronutrients and micronutrients at the stoichiometric ratios required by green algae and diatoms[35]. Each mesocosm was treated as a 'quasi-chemostat' in which 10% of the water volume was replaced with fresh medium each week to replenish nutrients and reduce waste.

Recirculating flumes are a widely used model system for laboratory studies of lotic algae and invertebrates[36]. In addition to the high level of replication that can be achieved, experiments can be run with large population sizes. For example, final population sizes on the growth substrates in this study averaged 1.2×10^9 cells for algae grown together in polyculture. For comparison, the entire Brazilian Amazon ($>4 \times 10^6\ km^2$) is estimated to have 2.7×10^{11} trees that are greater than 10-cm diameter breast height[37].

Experimental design. This experiment manipulated two variables in factorial combination: algal species richness (levels, one, two, four, six and eight species) × type of growth environment (levels, heterogeneous and homogeneous). Flumes assigned to the heterogeneous growth environment were constructed to have two forms of spatial and temporal variation that are widely thought to influence the diversity and coexistence of stream organisms (Supplementary Fig. 1). Spatial heterogeneity was added by angling the biofilm growth surface vertically in a flume, creating a flow constriction that generated spatial variation in near-bed velocities ranging from $<2\ cm\ s^{-1}$ at the bottom of the ramp to $55\ cm\ s^{-1}$ at the top of the ramp (velocities were measured by dissolution of gypsum pellets whose weight loss had been calibrated to free stream velocity in these flumes using a 16-MHz Micro Acoustic Doppler Velocimeter (SonTek)). The abundance, distribution and diversity of stream primary producers are known to be influenced by a stream's flow velocity[25,27,32,34,38], and species have evolved a variety of adaptations to deal with near-bed shear stress, as well as to sequester nutrients from boundary layers that control the delivery of materials to cells[26,27]. In comparison to the heterogeneous growth environments, growth substrates in the homogeneous flumes were laid flat on the bottom of the streams so that near-bed velocity could be set to a uniform $22\ cm\ s^{-1}$ (which is the median velocity of the heterogeneous environment).

To generate temporal variation in the heterogeneous flumes, growth substrates were divided into 18 equally sized patches, each of which had a probability of 0.25 of being 'disturbed' in each week of the experiment. When a patch was selected for disturbance, algae growing in that patch were removed using a soft-bristled bottle brush. This treatment was designed to generate a spatial mosaic of patches ranging in successional age from 5 to 50 days old, which is typical of streams that experience periodic disturbances owing to flow and sediment movement[28,39–42]. For flumes assigned to the homogeneous treatment, none of the 18 patches was disturbed, so all patches were uniformly 50 days old at the end of the experiment.

Treatments of algal diversity were applied as a randomized complete block in which biofilms differing in richness were established in 50 flumes (25 heterogeneous and 25 homogeneous) at a time, and the experiment was then repeated in three temporal blocks for a total of 150 mesocosms (75 per growth environment). In each block of the experiment, I grew each of the eight species alone as monocultures, five combinations of species at each intermediate level of richness (two, four and six species) with species selected randomly from all possible combinations, and two replicates of eight-species polycultures. Thus, for either type of environment (heterogeneous or homogeneous), the three experimental blocks included 3 replicate flumes for every species grown alone in monoculture, 15 replicate flumes at each level of two, four or six species, and 6 replicate full species polycultures.

At the start of each block, species were inoculated in the flumes according to a replacement series (that is, substitutive) design in which a total of 10,000 cells were added to the water irrespective of species richness. Flumes were inoculated on day 1, as well as day 7, of each block to ensure successful establishment. Each block lasted for 6 weeks. Pilot studies that were performed immediately before this study indicated that the 6-week time frame was sufficient to achieve ~10–20 doublings of algal population size (depending on the species) and for the full (eight-species) polycultures to reach a steady-state biomass.

Measurements. At the end of the experiment, I performed a ^{15}N tracer study to measure the rates of NO_3^- uptake by the biofilms. ^{15}N-enriched-$NaNO_3$ (60% ^{15}N) was added to sets of ten flumes at a time in intervals that were staggered by 20 min. The additions were sufficient to increase the $^{15}N/^{14}N$ ratio in the dissolved NO_3^- pool by a factor of 1.88 (1.65-, 2.00- and 1.99-fold for blocks 1–3). By contrast, additions resulted in a negligible change to the total concentration of NO_3^-, which increased by just 0.5% (1.004-, 1.006- and 1.006-fold for blocks 1–3). Thus, the $^{15}NO_3^-$ served as a tracer for nutrient uptake.

Biofilms were allowed to accumulate $^{15}NO_3^-$ for 2.4 ± 0.4 h. Then, the lights were turned off for sets of ten flumes at a time to prevent photosynthesis, and the biofilms in these flumes were destructively sampled over 20 min to obtain three measurements: final population size, final algal biomass and ^{15}N uptake.

To estimate final population sizes and to determine how these varied among habitat types in a flume, 0.49-cm^2 subsamples of the biofilm were taken from each of the 18 patches using a razor blade. The subsamples were preserved in 3% formalin and later used to identify species and to estimate cell densities using a haemocytometer viewed under a BX41 compound microscope (Olympus). The biofilm remaining on the 200-cm^2 growth surface (lacking the $8.8\ cm^2$ used for the 18 subsamples) was then removed with a razor blade and brought to a constant volume, and additional subsamples were taken for the other two measurements.

To estimate the final algal biomass in a flume, two subsamples of the biofilm were taken. The first was filtered onto a 0.45-μm GF-F filter (Whatman), which was sealed in a 10-ml Falcon tube containing 90% ethanol and placed in a freezer to lyse the cells and extract the photopigments. Concentrations of extracted chlorophyll *a*, which is a widely used estimate of algal biomass[14], were analysed spectrophotometrically using previously described methods[43], together with previously reported light extinction coefficients[44] for extracts in ethanol. A second subsample of the biofilm was dried for 48 h at 60 °C, after which total dry mass was determined. The Pearson correlation coefficient relating chlorophyll *a* and biofilm dry mass was 0.89, indicating that these two measures give nearly identical information. Chlorophyll *a* is reported on in this study, because it is specific to algae (that is, it does not include the mass of heterotrophs such as bacteria) and is perhaps the most widely used measure of algal biomass.

^{15}N uptake was estimated from a final subsample of the biofilm, which was dried for 48 h at 60 °C. The sample was then ground with a mortar and pestle, and preweighed samples were packed into combustible tin capsules. All mass spectrometry was performed at the University of California Santa Barbara's Marine Science Institute analytical laboratory. Samples were analysed using a DeltaPLUS isotope ratio mass spectrometer (Finnigan). ^{15}N values were expressed as $\delta^{15}N$, calculated as $\delta^{15}N = [(R_{sample}/R_{standard}) - 1] \times 1{,}000$, where $R = {^{15}N}/{^{14}N}$ ratio and $R_{standard}$ is the $^{15}N/^{14}N$ ratio in atmospheric N_2 (0.0036765). Del values were converted to atom per cent, *AP*, as given in the appendix of ref. 45: $AP = 100(\delta + 1{,}000)/[(\delta + 1{,}000 + (1{,}000/R_{standard})]$. Nitrogen uptake rates were then calculated using equation 6 from ref. 46: $N_{uptake}\ (\mu g\ cm^{-2}\ h^{-1}) = [(AP_{is} - AP_{ns})/(AP_{ic} - AP_{ns})](N/t)$, where AP_{is} is the atom per cent of ^{15}N in the biofilm after incubation, AP_{ns} is the atom per cent of ^{15}N in algae before incubation (determined by analysis of three replicate samples from each algal batch culture used to inoculate the flumes), AP_{ic} is the atom per cent of ^{15}N in the dissolved phase at the beginning of the incubation (which was assumed to be in equilibrium

©2011 Macmillan Publishers Limited. All rights reserved

with air), N is the total nitrogen in the biofilm after incubation ($\mu g\,cm^{-2}$) and t is the incubation time in h.

Data analyses. N_{uptake} rates and algal biomass were both modelled as a function of algal species richness using three functions that have commonly been used to describe relationships between biological diversity and ecological processes in the literature[22]. These include a linear function, a positive but decelerating function (log) and a positive saturating function (the Michaelis–Menten hyperbolic equation). Parameter estimates were generated by maximum likelihood, using mixed model analyses that included blocking as a random effect. Akaike information criteria were used to judge the relative fits of the functions to the data[47].

It is methodologically impossible to distinguish how much ^{15}N or chlorophyll *a* is incorporated into the cells of different species grown together in a polyculture, as there is no practical way to separate the species physically. However, it is possible to distinguish between two factors that contributed to the final population densities of species in polyculture using an approach developed in ref. 24. With this formula, any difference in the biomass or density of two levels of species richness, ΔY, can be statistically partitioned into two additive components:

$$\Delta Y = N\overline{\Delta RY}\bar{M} + [N cov(\Delta RY, M)] \qquad (1)$$

where N is the number of species in a polyculture, RY is the relative yield (which can be based on biomass or density) of species in a polyculture and M is the yield of species in a monoculture. The notation ΔRY refers to the deviation in relative yield from some expected value, which is usually taken to be the proportional inoculation density of a species. For a replacement-series design such as that used here, $\Delta RY = I/N$, where I is the total inoculated density and N is the number of species in polyculture. In equation (1), the first term on the right-hand side measures what is commonly referred to as the complementarity effect (CE). Complementarity is positive whenever species yields in a mixture are, on average, higher than those that would be expected from the weighted average yield of the species in monoculture. This can occur when species use habitats or resources in ways that are complementary to one another, such as through niche partitioning, facilitation or other interactions that enhance species population sizes and/or per capita growth[48,49]. The second term on the right-hand side of equation (1) measures what is referred to as the selection effect (SE). Positive selection occurs if species with higher-than-average monoculture yields competitively dominate the polycultures.

31. Biggs, B. J. F., Goring, D. G. & Nikora, V. I. Subsidy and stress responses of stream periphyton to gradients in water velocity as a function of community growth form. *J. Phycol.* **34,** 598–607 (1998).
32. Passy, S. I. Spatial paradigms of lotic diatom distribution: a landscape ecology perspective. *J. Phycol.* **37,** 370–378 (2001).
33. Biggs, B. J. F., Stevenson, R. J. & Lowe, R. L. A habitat matrix conceptual model for stream periphyton. *Arch. Hydrobiol.* **143,** 21–56 (1998).
34. Stevenson, R. J. Effects of current and conditions simulating autogenically changing microhabitats on benthic diatom immigration. *Ecology* **64,** 1514–1524 (1983).
35. Andersen, R. A. *Algal Culturing Techniques* (Elsevier/Academic, 2005).
36. Vogel, S. & LaBarbera, M. Simple flow tanks for research and teaching. *Bioscience* **28,** 638–645 (1978).
37. Hubbell, S. P. *et al.* How many tree species are there in the Amazon and how many of them will go extinct? *Proc. Natl Acad. Sci. USA* **105,** 11498–11504 (2008).
38. Stevenson, R. J. in *Algal Ecology: Freshwater Benthic Ecosystems* (eds Stevenson, R. J., Bothwell, M. L. & Lowe, R. L.) 321–336 (Academic, 1996).
39. Poff, N. L. *et al.* The natural flow regime. *Bioscience* **47,** 769–784 (1997).
40. Townsend, C. R. *et al.* Disturbance, resource supply, and food-web architecture in streams. *Ecol. Lett.* **1,** 200–209 (1998).
41. Cooper, S., Barmuta, L., Sarnelle, O., Kratz, K. & Diehl, S. Quantifying spatial heterogeneity in streams. *J. N. Am. Benthol. Soc.* **16,** 174–188 (1997).
42. Townsend, C. R. The patch dynamics concept of stream community ecology. *J. N. Am. Benthol. Soc.* **8,** 36–50 (1989).
43. Steinman, A. D. & Lamberti, G. A. in *Methods in Stream Ecology* (eds Hauer, F. R. & Lamberti, G. A.) 295–311 (Academic, 1996).
44. Nusch, E. A. Comparison of different methods for chlorophyll and phaeopigment determination. *Arch. Hydrobiol. Beih. Ergebn. Limnol.* **14,** 14–36 (1980).
45. Fry, B. *Stable Isotope Ecology* (Springer, 2006).
46. Legendre, L. & Gosselin, M. Estimation of N or C uptake rates by phytoplankton using N-15 or C-13: revisiting the usual computation formulae. *J. Plankton Res.* **19,** 263–271 (1997).
47. Johnson, J. B. & Omland, K. S. Model selection in ecology and evolution. *Trends Ecol. Evol.* **19,** 101–108 (2004).
48. Carroll, I. T., Cardinale, B. J. & Nisbet, R. M. Niche and fitness differences relate the maintenance of diversity to ecosystem function. *Ecology* (in the press).
49. Loreau, M. Does functional redundancy exist? *Oikos* **104,** 606–611 (2004).

©2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09807

Tumour evolution inferred by single-cell sequencing

Nicholas Navin[1,2], Jude Kendall[1], Jennifer Troge[1], Peter Andrews[1], Linda Rodgers[1], Jeanne McIndoo[1], Kerry Cook[1], Asya Stepansky[1], Dan Levy[1], Diane Esposito[1], Lakshmi Muthuswamy[3], Alex Krasnitz[1], W. Richard McCombie[1], James Hicks[1] & Michael Wigler[1]

Genomic analysis provides insights into the role of copy number variation in disease, but most methods are not designed to resolve mixed populations of cells. In tumours, where genetic heterogeneity is common[1–3], very important information may be lost that would be useful for reconstructing evolutionary history. Here we show that with flow-sorted nuclei, whole genome amplification and next generation sequencing we can accurately quantify genomic copy number within an individual nucleus. We apply single-nucleus sequencing to investigate tumour population structure and evolution in two human breast cancer cases. Analysis of 100 single cells from a polygenomic tumour revealed three distinct clonal subpopulations that probably represent sequential clonal expansions. Additional analysis of 100 single cells from a monogenomic primary tumour and its liver metastasis indicated that a single clonal expansion formed the primary tumour and seeded the metastasis. In both primary tumours, we also identified an unexpectedly abundant subpopulation of genetically diverse 'pseudodiploid' cells that do not travel to the metastatic site. In contrast to gradual models of tumour progression, our data indicate that tumours grow by punctuated clonal expansions with few persistent intermediates.

In single-nucleus sequencing (SNS), we isolate nuclei by flow-sorting and amplify DNA using whole genome amplification (WGA) for massively parallel sequencing (Supplementary Fig. 1). We achieve low coverage (~6%) of the genome of a single cell, sufficient to quantify copy number from sequence read depth. Several features of our data analysis were designed for SNS and differ from previous methods[4–6] for measuring copy number from sequencing data. In contrast to using fixed intervals to calculate copy number, we use variable length bins but with uniform expected unique counts, which correct for biases that have been reported[7–9] in WGA (Supplementary Fig. 2; see Methods). For each single cell, we typically achieve a mean read density of 138 per bin (standard error of the mean (s.e.m.) ± 5.55, $n = 200$). Over-replicated loci called 'pileups', which have been previously reported in WGA[10–12], do occur in our data but not at recurrent locations in different cells (Supplementary Fig. 3). Pileups are sufficiently randomly distributed and sparse so as not to affect counting at the resolution we

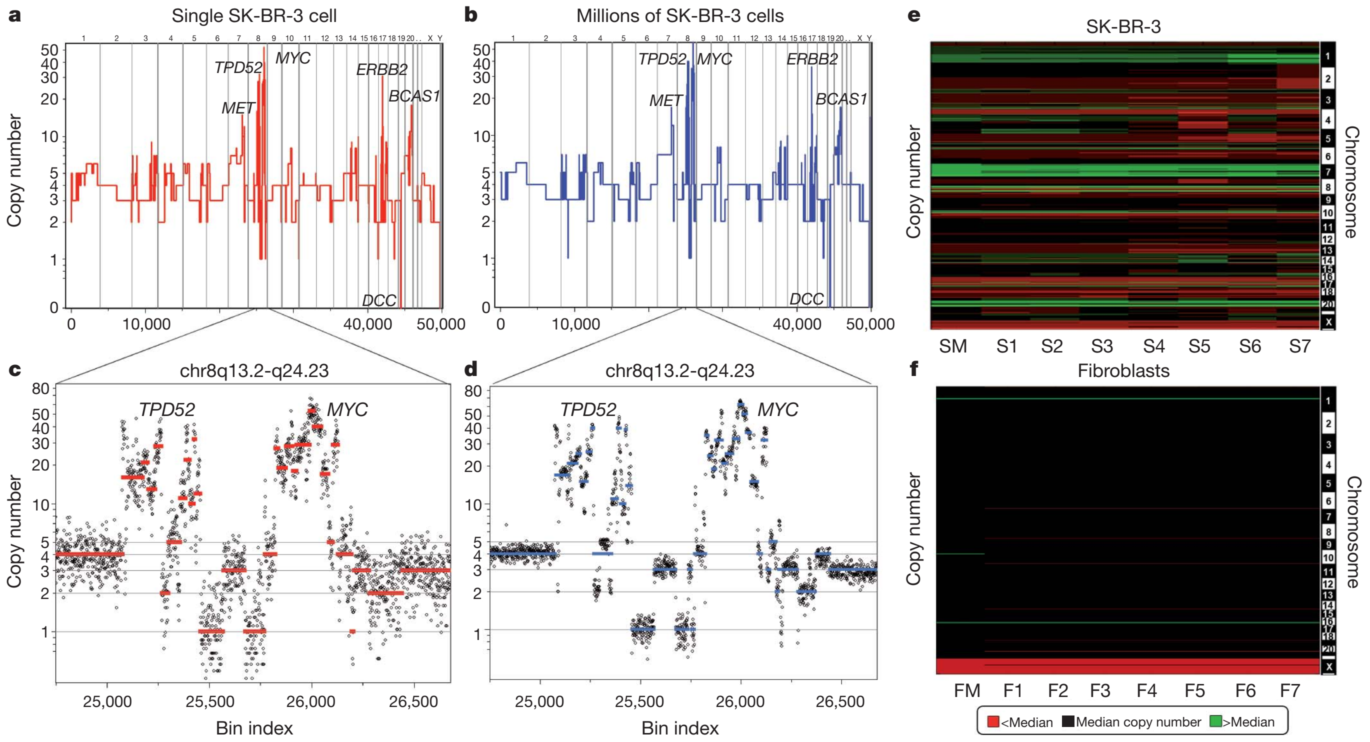


Figure 1 | Comparison of SK-BR-3 single cells to millions. **a**, **b**, The integer copy number profile for a single SK-BR-3 cell is shown (**a**) compared to a sequence count profile using millions of cells (**b**). **c**, **d**, A region on chromosome 8q13.2-q24.23 is plotted showing the integer copy number profile (in red or blue) and a ratio of raw bin counts in grey for a single cell (**c**), and a million cells (**d**). **e**, A heatmap of SK-BR-3 copy number profiles comparing a million-cell sample (SM) to seven single cells (S1–S7). **f**, A heatmap of SKN1 normal fibroblast profiles comparing a million-cell sample (FM) to seven single cells (F1–F7).

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. [2]Department of Genetics, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [3]Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada.

©2011 Macmillan Publishers Limited. All rights reserved

have chosen (54 kb). Assuming that single cells will have discrete copy number states, we segment the variable bins and calculate integer copy number profiles (Supplementary Fig. 4; see Methods).

To validate our method, we compared the sequence counting profile of DNA from a single SK-BR-3 cell (Fig. 1a) with DNA from one million cells (Fig. 1b). The major amplifications (*MET*, *TPD52*, *ERBB2*, *BCAS1*) and deletions (*DCC*) are detected in both profiles, as are much more abundant but less marked small changes in copy number. To demonstrate how reproducible small differences are, we assessed data for a complex region on chromosome 8q13.2-q24.23 that contains more than thirty segments with differing copy number. These data were reproducible in both a single-cell (Fig. 1c) and a million-cell sample (Fig. 1d). We also compared the sequence read profiles from several single cells and from a million cells to each other and to the profile measured by microarray comparative genomic hybridization (CGH) from bulk DNA (Supplementary Fig. 5). In all instances the profiles showed very high ($r^2 > 0.85$) correlation. The reproducibility and variation between single-cell copy number profiles was also investigated by comparing seven single cells from a culture of SK-BR-3 and seven from normal human fibroblasts. These data are shown as heat maps (Fig. 1e–f), which show that some genomic variation exists between cells. The diploid fibroblast cultures showed no random events; we observed only a few consistent events at levels expected for heritable copy number variations.

We selected next two high-grade (III), triple-negative (ER$^-$, PR$^-$, HER2$^-$) ductal carcinomas (T10, T16P) and a paired metastatic liver carcinoma (T16M) to study tumour population structure and infer tumour evolution by single-cell analysis. T10 was selected to study primary tumour growth because it was previously shown[13] to be genetically heterogeneous (polygenomic), and T16P was selected because it was classified as genetically homogeneous (monogenomic).

T10 was macrodissected into 12 sectors to preserve anatomical information, and nuclei were flow-sorted from six sectors (S1–S6) for SNS (Fig. 2a). Fluorescence-activated cell sorting (FACS) analysis

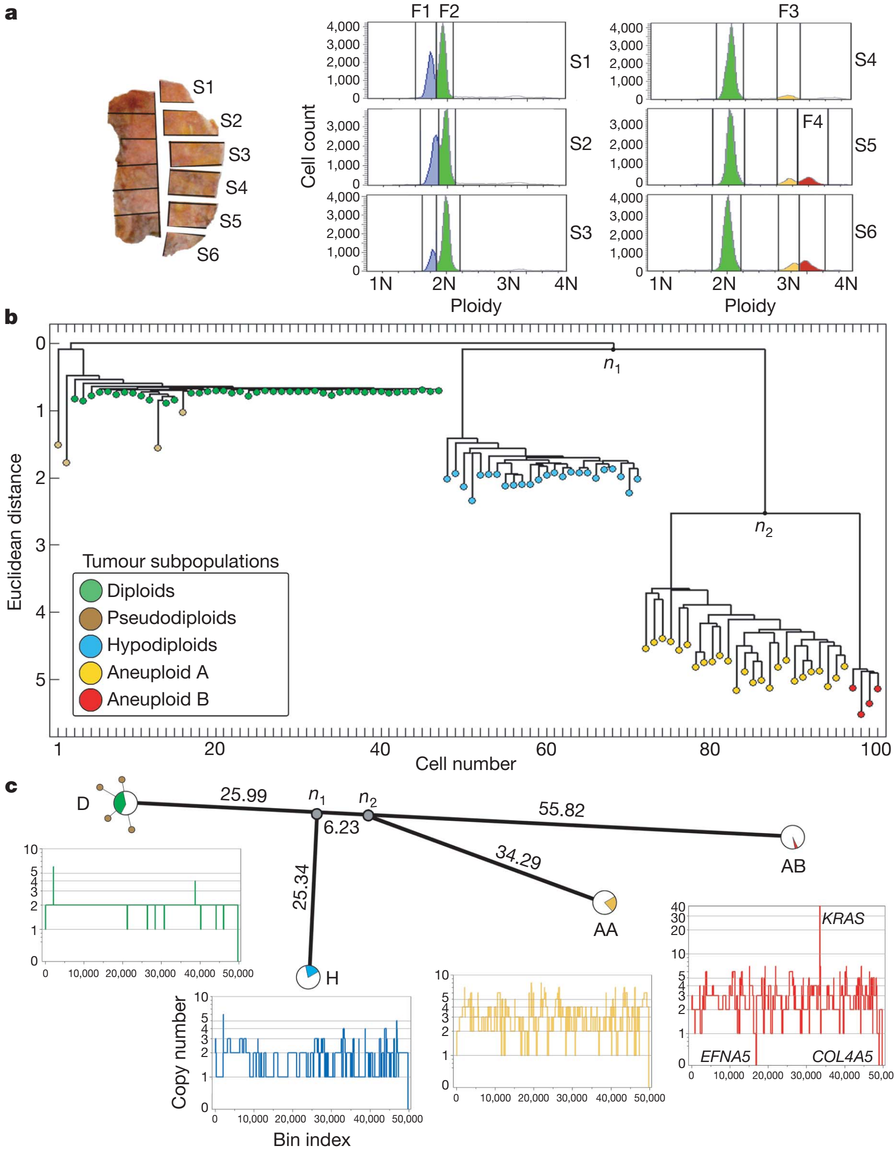


Figure 2 | Analysis of 100 single cells from a polygenomic breast tumour. **a**, T10 was macrodissected into 12 sectors, and nuclei were isolated from six sectors and flow-sorted by ploidy. FACS profiles show four distributions of ploidy (F1–F4), which were gated to isolate 100 single cells. **b**, Neighbour-joining tree of integer copy number profiles showing four major branches of evolution. **c**, Phylogenetic tree of consensus profiles show the common ancestors and evolutionary distance between subpopulations. Integer copy number profiles from single cells are displayed below, and pie charts indicate the percentage of cells that constitute each subpopulation.

©2011 Macmillan Publishers Limited. All rights reserved

showed four major distributions of ploidy: a hypodiploid fraction (F1) exclusive to sectors 1–3; a diploid 2N fraction (F2) in all sectors; and two subtetraploid fractions (F3 and F4) in sectors 4–6. We selected 100 single cells from multiple sectors and ploidy fractions for sequencing and calculation of integer copy number profiles (Supplementary Table 1).

Breast tumours are typically mixtures of cancer cells with normal tissue, stroma and infiltrating leukocytes. By histopathology, T10 was assessed to contain 63% normal and 37% tumour cells and noted to be heavily infiltrated with leukocytes. Most of the diploid nuclei from F2 had flat genome profiles, characteristic of normal cells. Nearly two-thirds (31/47) of these diploid profiles showed narrow deletions in the T-cell receptor loci or one or more immunoglobulin variable region loci, consistent with infiltration by immunocytes (data not shown). Of the remaining sixteen nuclei from F2, twelve showed no discernable aberrations, but four nuclei showed aberrant profiles with diverse chromosome gains and losses. Each of these 'pseudodiploid' nuclei profiles seemed unrelated to the others or to those of the major tumour cell populations found in fractions F1, F3 and F4.

To determine population substructure we calculated pair-wise distances between the 100 integer copy number profiles, and built a tree using neighbour joining[14] (Fig. 2b). The 100 profiles clustered into four subpopulations (D+P, H, AA and AB) regardless of their sector of origin. The D+P subpopulation contains predominantly flat diploid (D) profiles, but also pseudodiploid (P) cells that have diverged by varying degrees from the diploids. The three major 'advanced' tumour subpopulations (H, AA and AB) are highly clonal with complex genomic rearrangements, and together comprise slightly less than half the cells of the tumour. These cells were isolated from the hypodiploid (F1) and two subtetraploid (F3 and F4) ploidy fractions, respectively. We had previously identified these subpopulations by profiling millions of cells by array CGH[13], but we could not determine if they were composite mixtures of different tumour clones. By SNS we can now see that each subpopulation is composed of cells that share highly similar copy number profiles, probably representing three clonal expansions. Each subpopulation (H, AA and AB) is clearly related to the others by many shared genomic alterations, but they have also diverged and developed distinct attributes (for example, a massive 50-fold amplification of the *KRAS* oncogene in AB). The H cells display the characteristic 'sawtooth' pattern[15] comprising broad chromosomal deletions (Fig. 2c). They are anatomically segregated in sectors S1–S3 of the tumour, whereas the AA and AB clones are intermixed and occupy sectors S4–S6.

To understand the relationship between subpopulations, we clustered profiles by chromosome breakpoints (which are directly related to the steps by which tumour cells diverge). We identified 657 copy number breakpoints and used them to build a phylogenetic tree, which closely resembles the structure of the neighbour-joining tree based on copy number (Supplementary Fig. 6). We also applied biclustering[16] to construct a heat map of breakpoints, and ordered it on the basis of the copy number tree to show which breakpoints were common or divergent between the major subpopulations (Supplementary Fig. 7a). Although there is considerable variation within each subpopulation, no obvious further population substructure was evident. To estimate the common ancestors, we constructed a phylogenetic lineage using the consensus breakpoint patterns from the

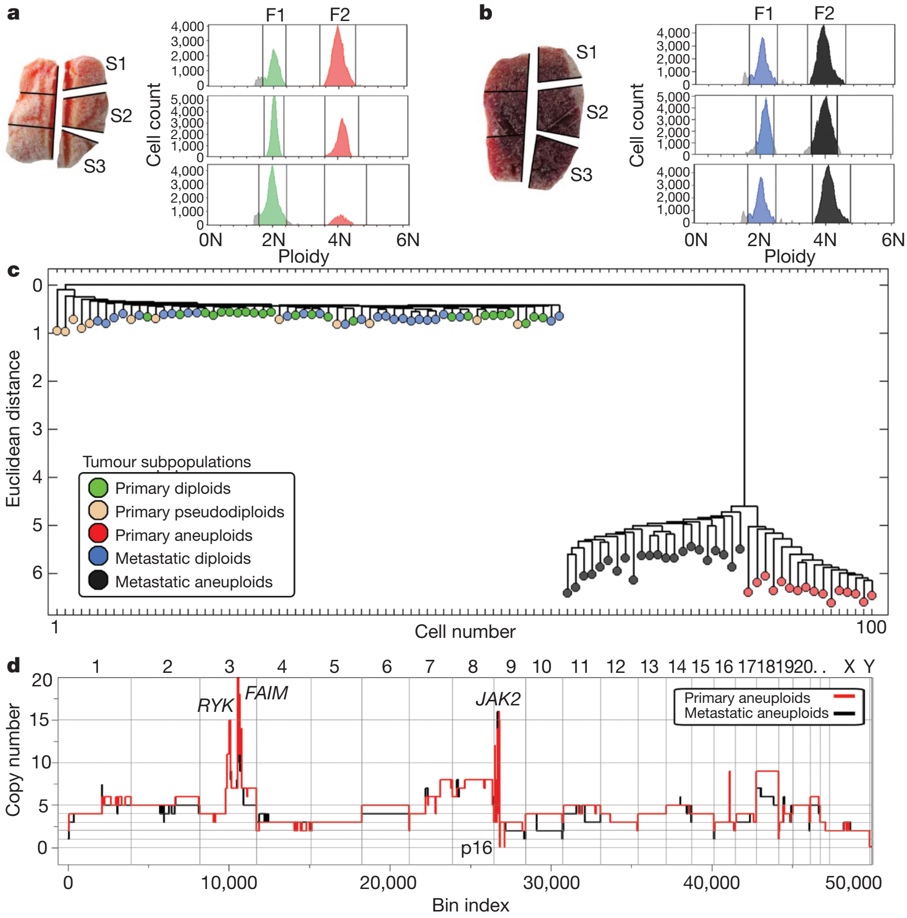


Figure 3 | Analysis of 100 single cells from a monogenomic breast tumour and its liver metastasis. a, **b**, Primary breast tumour T16P was macrodissected and 52 nuclei were isolated from three sectors for FACS, showing two distributions of ploidy (F1 and F2). **b**, Liver metastasis T16M was macrodissected and 48 nuclei were isolated from three sectors for FACS also showing two ploidy distributions (F1 and F2). **c**, Neighbour-joining tree of combined integer copy number profiles from the primary and metastatic tumours. **d**, Comparison of primary and metastatic aneuploid consensus copy number profiles.

©2011 Macmillan Publishers Limited. All rights reserved

major tumour subpopulations (Fig. 2c). This lineage shows that the n_1 common ancestor diverged a significant distance from the diploid cells, but that the distance between n_1 and n_2 is very small. By contrast, the divergence of the subpopulations after n_1 and n_2 is very large, with AB showing the greatest phylogenetic distance from the diploids. Thus we infer that the three subpopulations emerged when the tumour was much smaller.

We investigated a second tumour to determine whether these findings extend. We isolated 52 cells from a primary breast tumour (T16P) and 48 cells from its associated liver metastasis (T16M). Each tumour was macrodissected into six sectors, three of which were flow-sorted (Fig. 3a, b). Both T16M and T16P showed diploid peaks (F1) and a single aneuploid tetraploid peak (F2) of roughly equal cell count in all sectors (Supplementary Table 2), consistent with histological sections showing approximately 50% tumour and 50% normal (stromal) cells with low leukocyte infiltration in both samples. To explore population substructure we again constructed neighbour-joining trees from the integer copy number profiles, combining the primary and metastasis cells (Fig. 3c). We observed again numerous pseudodiploid cells, but a single subpopulation of aneuploid cells very diverged from the diploid population. As for T10, the 12 pseudodiploid cells from T16P showed diverse genomic lesions with no clear relationships to each other or to the main tumour lineage. Of the 24 normal diploids in the primary, two had deletions of the T-cell receptor. There were no pseudodiploid cells among the 26 diploid cells from the metastasis.

These data indicate that the primary tumour mass formed by a single clonal expansion of an aneuploid cell, and that one of the cells from this expansion subsequently seeded the metastatic tumour with little further evolution. There are no branches of the tree corresponding to cells intermediate between the aneuploid subpopulation and the diploid root. Although closely related, the primary and metastatic aneuploid cells cleanly separate using the Euclidean metric (Fig. 3c), indicating that the two populations have not mixed since seeding the metastasis. The differences in the profiles that distinguish the primary and metastatic tumour populations are in the degree of copy number change rather than breakpoints (Fig. 3d). In a hierarchical tree created from breakpoints alone, we cannot cleanly separate primary from metastatic aneuploid cells (Supplementary Fig. 6b). Moreover, when we calculate common breakpoints in the single-cell profiles and apply biclustering to ordered samples (Supplementary Fig. 7b), a large number of breakpoints are common to both populations and no breakpoints cleanly distinguish them. By these analyses, no further population substructure is evident.

In contrast to the clear clonal relationships among aneuploid subpopulations, pseudodiploid cells are unusual in showing remarkable genomic heterogeneity (Fig. 4). Pseudodiploid profiles are characterized by nonrecurring copy number changes (including whole chromosome arms) that are not shared between any two pseudodiploid cells, nor with the corresponding tumour profiles (Fig. 4e). These data indicate that unlike the aneuploid cells, pseudodiploids do not undergo clonal expansions in the tumour. Nevertheless, they comprise a substantial proportion of the diploid gated cells: 8% in T10 (4/47) and 33% in T16P (12/36), or approximately 4% and 24% of the tumour mass, respectively. In contrast, the 18 profiles from single nuclei of normal adjacent breast tissue are all flat (Fig. 4a). The relative abundance of pseudodiploid cells in primary tumours indicates that they may emerge from an ongoing aberrant process that generates genomic diversity in the tumour.

In principle, we can learn about DNA sequence mutations from SNS data. However, the sparse sequence coverage makes this analysis problematic. By combining data from multiple cells, belonging to well-defined subpopulations, we can perform global and regional analysis at the many nucleotide positions where sufficient numbers of sequence

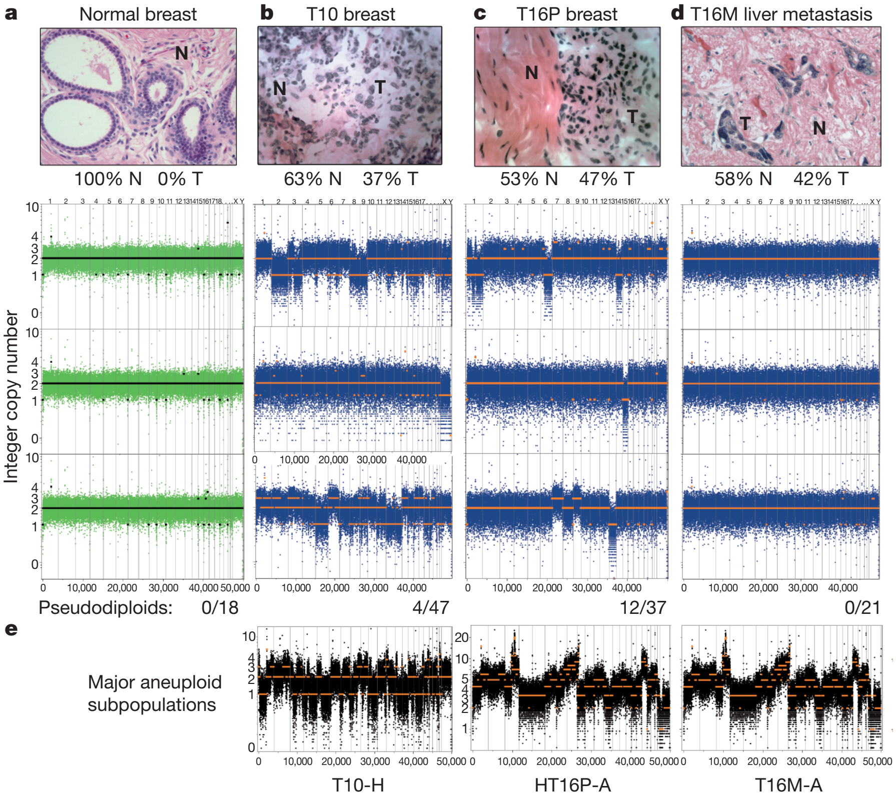


Figure 4 | Genetically diverse pseudodiploid cells in the diploid fractions of tumours. **a**–**d**, Haematoxylin and eosin stained tissues sections are shown in the upper panels with normal (N) and tumour (T) cell percentages indicated. Lower rows show bin counts and copy number profiles of single cells isolated from the 2N gated ploidy distributions, and the total number of cells analysed is indicated below each column. The columns are: normal breast tissue cells (**a**); pseudodiploid cells in T10 (**b**); pseudodiploid cells in T16P (**c**); and diploid-gated nuclei from T16M (**d**). **e**, Bin counts and copy number profiles of single cells from the major aneuploid tumour subpopulations.

©2011 Macmillan Publishers Limited. All rights reserved

reads overlap. When examined this way, losses of heterozygosity are unequivocally significant, and map in large contiguous genomic blocks that correlate well with copy number loss (Supplementary Fig. 8 and Supplementary Table 3). The extensive loss of heterozygosity detected in all of the T10 subpopulations and in T16 indicates that both cancers passed through a hypodiploid stage.

Our study demonstrates that we can obtain robust high-resolution copy number profiles by sequencing a single cell and that by examining multiple cells from the same cancer we can make inferences about the evolution and spread of cancer. Moreover, the identification of pseudodiploid cells shows that these methods can identify cell types previously undetectable by other methods. Our findings are consistent with previous findings[17] using bulk DNA, which indicate that copy number profiles in primary tumours are highly similar to the metastases. Thus, the metastatic cells emerge from a main advanced expansion, and not from an earlier intermediate or a completely different subpopulation. This is consistent with recent deep-sequencing studies of primary–metastatic pairs, all indicating that metastatic cells arise late in tumour development[18,19].

There are many gradual models for tumour progression, including clonal evolution[20], the mutator phenotype[21,22] and stochastic progression[23]. Although we have examined only two cancers in depth, both show a pattern of tumour growth that we call 'punctuated clonal evolution', borrowing a term from species evolution used to explain gaps in the fossil record[24]. Explicitly, the tumour subpopulations are each distant from their root, without observable intermediate branching. In contrast to gradual models, this pattern reflects the sudden emergence of a tumour cell whose rate of effective population growth markedly exceeds its rate of genomic evolution.

METHODS SUMMARY

To perform SNS, nuclei are isolated either from cells in culture or frozen tumour sections and stained with 4′,6-diamidino-2-phenylindole (DAPI). We use FACS to gate a desired population of nuclei by total DNA content and to deposit nuclei singly into 96-well plates. After WGA using Sigma GenomePlex, we sonicate to create free DNA ends without WGA adapters, and then construct libraries for 76 bp, single-end sequencing using one lane of an Illumina GA2 flowcell per nucleus. For each nucleus we typically achieve 9 million (mean = 9.042 million, s.e.m. ± 0.328, $n = 200$) uniquely mapping reads using the Bowtie[25] alignment software. These sequences cover about 6% (mean = 5.95%, s.e.m. ± 0.229, $n = 200$) of the genome, and are used to count sequence reads in 50,000 variable bins. The bin counts are segmented using a KS statistic and used to calculate integer copy number profiles. Neighbour-joining trees are constructed from the integer profiles and from the chromosome breakpoint patterns of each cell to infer evolution.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 May 2010; accepted 7 January 2011.
Published online 13 March 2011.

1. Park, S. Y., Gonen, M., Kim, H. J., Michor, F. & Polyak, K. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *J. Clin. Invest.* **120,** 636–644 (2010).
2. Torres, L. *et al.* Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. *Breast Cancer Res. Treat.* **102,** 143–155 (2007).
3. Farabegoli, F. *et al.* Clone heterogeneity in diploid and aneuploid breast carcinomas as detected by FISH. *Cytometry* **46,** 50–56 (2001).
4. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods* **6,** 99–103 (2009).
5. Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* **19,** 1586–1592 (2009).
6. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genet.* **41,** 1061–1067 (2009).
7. Geigl, J. B. *et al.* Identification of small gains and losses in single cells after whole genome amplification on tiling oligo arrays. *Nucleic Acids Res.* **37,** e105 (2009).
8. Fuhrmann, C. *et al.* High-resolution array comparative genomic hybridization of single micrometastatic tumor cells. *Nucleic Acids Res.* **36,** e39 (2008).
9. Pugh, T. J. *et al.* Impact of whole genome amplification on analysis of copy number variants. *Nucleic Acids Res.* **36,** e80 (2008).
10. Talseth-Palmer, B. A., Bowden, N. A., Hill, A., Meldrum, C. & Scott, R. J. Whole genome amplification and its impact on CGH array profiles. *BMC Res. Notes* **1,** 56 (2008).
11. Hughes, S. *et al.* Use of whole genome amplification and comparative genomic hybridisation to detect chromosomal copy number alterations in cell line material and tumour tissue. *Cytogenet. Genome Res.* **105,** 18–24 (2004).
12. Huang, J., Pang, J., Watanabe, T., Ng, H. K. & Ohgaki, H. Whole genome amplification for array comparative genomic hybridization using DNA extracted from formalin-fixed, paraffin-embedded histological sections. *J. Mol. Diagn.* **11,** 109–116 (2009).
13. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20,** 68–80 (2010).
14. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).
15. Hicks, J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* **16,** 1465–1479 (2006).
16. Prelic, A. *et al.* A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **22,** 1122–1129 (2006).
17. Liu, W. *et al.* Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nature Med.* **15,** 559–565 (2009).
18. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).
19. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467,** 1114–1117 (2010).
20. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
21. Loeb, L. A., Springgate, C. F. & Battula, N. Errors in DNA replication as a basis of malignant changes. *Cancer Res.* **34,** 2311–2321 (1974).
22. Bielas, J. H., Loeb, K. R., Rubin, B. P., True, L. D. & Loeb, L. A. Human cancers express a mutator phenotype. *Proc. Natl Acad. Sci. USA* **103,** 18238–18242 (2006).
23. Heng, H. H. *et al.* Stochastic cancer progression driven by non-clonal chromosome aberrations. *J. Cell. Physiol.* **208,** 461–472 (2006).
24. Gould, S. J. & Eldredge, N. Punctuated equilibria comes of age. *Nature* **366,** 223–227 (1993).
25. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Ronemus, T. Spencer, A. Leotta, J. Meth, M. Kramer, L. Gelley, E. Ghiban. We also thank P. Blake and N. Navin at Sophic Systems Alliance. This work was supported by the NCI T32 Fellowship to N.N., and grants to M.W. and J.H. from the Department of the Army (W81XWH04-1-0477), the Breast Cancer Research Foundation, and the Simons Foundation. M.W. is an American Cancer Society Research Professor.

Author Contributions N.N. designed and performed experiments and analysis, and wrote the manuscript. J.K., A.K., L.M., D.L. and P.A. developed analysis programs. J.T., L.R., K.C., J.M., D.E. and A.S. performed experiments. W.R.M. designed experiments. J.H. and M.W. designed experiments, performed analysis and wrote manuscript.

Author Information All data has been deposited into the NCBI Sequence Read Archive under accession number SRA018951.105. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.W. (wigler@cshl.edu).

©2011 Macmillan Publishers Limited. All rights reserved

METHODS

Samples. The frozen ductal carcinoma T10 (CHTN0173) was obtained from the Cooperative Human Tissue Network, and T16P and T16M were obtained from Asterand. Pathology shows that both tumours were poorly differentiated and high grade (III) as determined by the Bloom–Richardson score, and triple-negative (ER^-, PR^- and $HER2/NEU^-$) as determined by immunohistochemistry. The cell lines used in this study include a normal male immortalized skin fibroblast (SKN1) and a breast cancer cell line (SK-BR-3). Normal breast tissue was obtained from H. Hibshoosh from Columbia University.

SNS. Nuclei were isolated from cell lines and from the frozen tumour using an NST-DAPI buffer (800 ml of NST (146 mM NaCl, 10 mM Tris base at pH 7.8, 1 mM $CaCl_2$, 21 mM $MgCl_2$, 0.05% BSA, 0.2% Nonidet P-40)), 200 ml of 106 mM $MgCl_2$, 10 mg of DAPI, and 0.1% DNase-free RNase A. The frozen tumour was first macrodissected into 12 sectors of equal size using surgical scalpels and nuclei were isolated from six sectors for FACS by finely mincing a tumour sector in a Petri dish in 1.0–2.0 ml of NST-DAPI buffer using two no. 11 scalpels in a cross-hatching motion. The cell lines were lysed directly in a culture plate using the NST-DAPI buffer, after first removing the cell culture media. All nuclei suspensions were filtered through 37-µm plastic mesh before flow-sorting.

Single nuclei were sorted by FACS using the BD Biosystems Aria II flow cytometer by gating cellular distributions with differences in their total genomic DNA content (or ploidy) according to DAPI intensity. First, a small amount of prepared nuclei from each tumour sample was mixed with a diploid control sample (derived from a lymphoblastoid cell line of a normal person) to accurately determine the diploid peak position within the tumour and establish FACS collection gates. Before sorting single nuclei, a few thousand cells were sorted to determine the DNA content distributions for gating. A 96-well plate was prepared with 10 µl of lysis solution in each well from the Sigma-Aldrich GenomePlex WGA4 kit. Single nuclei were deposited into individual wells in the 96-well plate along with several negative controls in which no nuclei were deposited.

WGA was performed on single flow-sorted nuclei as described in the Sigma-Aldrich GenomePlex WGA4 kit (catalogue no. WGA4-50RXN) protocol. WGA fragments from the frozen breast tumour and SK-BR-3 single cells were used directly for single-read library construction using the Illumina Genomic DNA Sample Prep Kit (catalogue no. FC-102-1001) and following standard protocol with a gel purification size range of 300–250 bp. WGA fragments from the fibroblast cell line were first sonicated using the Diagenode Bioruptor using the following program: 2 times, 7 min with 30 s high on/off mode in ice-cold water. Sonication removes a specific 28 bp adaptor sequence that is added on during WGA, and improves the total number of sequencing reads per lane.

Single-read libraries from single nuclei were sequenced on individual flow-cell lanes using the Illumina GA2 analyser for 76 cycles. Data was processed using the Illumina GAPipeline-1.3.2 to 1.6.0. Sequence reads were aligned to the human genome (HG18/NCBI36) using the Bowtie alignment software[25] with the following parameters: 'bowtie –S –t –m 1 –best –strata –p16' to report only top scoring unique mappings for each sequence read. For each nucleus we typically achieve 9 million (mean = 9.042 million, s.e.m. ± 0.328, $n = 200$) uniquely mapping reads. These sequences cover about 6% (mean = 5.95%, s.e.m. ± 0.229, $n = 200$) of the genome uniquely. To eliminate PCR duplicates, we removed sequences with identical start coordinates.

Read depth counting in variable bins. Copy number is calculated from read density, by dividing the genome into 'bins' and counting the number of unique reads in each bin. In previous copy number studies read density was calculated using bins with uniform fixed length[16–19]. In contrast, we use bins of variable length that adjust size depending on the mappability of sequences to regions of the human genome. In regions of repetitive elements, lower numbers of reads are expected and thus the bin size is increased. To determine interval sizes we simulated sequence reads by sampling 200 million sequences of length 48 from the human reference genome (HG18/NCBI36) and introduced single nucleotide errors with a frequency encountered during Illumina sequencing. These sequences were mapped back to the human reference genome using Bowtie[25] with unique parameters as described earlier. We assigned a number of bins to each chromosome based on the proportion of simulated reads mapped. We then divided each chromosome into bins with an equal number of simulated reads. This resulted in 50,009 genomic bins with no bins crossing chromosome boundaries. The median genomic length spanned by each bin is 54 kb. For each cell the number of reads mapped to each variable length bin was counted. This variable binning efficiently reduces false deletion events when compared to uniform length-fixed bins as shown in Supplementary Fig. 2b and c. For a single cell we typically measure 138 sequence reads per bin.

Integer copy number quantification. Single cells will have integer copy number states that we can infer from sequence read counts, as follows. Unique sequence reads are counted in variable bins (Supplementary Fig. 4a) and segmented using the Kolmogorov–Smirnov (KS) statistic (Supplementary Fig. 4b). To estimate the integer differences of copy number states, we calculate Gaussian kernel smoothed density plots using Splus (MathSoft), showing the difference between median bin counts for all pair-wise combinations of different segments (Supplementary Fig. 4c–e). The uniform steps between groups are very apparent, and are a general property of single-cell data. We then convert our KS-segmented data into profiles of integer copy number as follows. We take the differential bin count of the second peak, denoted by an asterisk in Supplementary Fig. 4a, to represent a copy number 'increment' of 1. We then divide every bin count in the profile by the increment and round to infer the integer copy number. We show in Supplementary Fig. 4f–g how closely the segmentation profile agrees with the integer copy number profile. However, for diploid or near diploid cells there are few to no steps from which to observe the increment, and we use a different method, taking the increment as the median bin count on the autosomes divided by two.

Gene annotations. Amplifications and deletions identified in the single-cell copy number profiles were annotated to identify UCSC genes. Cancer genes were identified using a compiled database from the cancer gene consensus and the NCI cancer gene index (Sophic Systems Alliance, Biomax Informatics AG).

Neighbour-joining trees of copy number profiles. Integer copy number profiles of single cells were used to calculate neighbour-joining trees using a Euclidean distance metric with Matlab (Mathworks). Branches were flipped to orient nodes within subpopulations and trees were rooted using the last common diploid node.

Common breakpoint detection. Breakpoints are defined as bins with a copy number different than the previous bin in genome order. A transition from a lower copy number to a higher copy number (in genome order) is considered to be a different event than the opposite transition. To find breakpoint regions we count each breakpoint in each cell and the immediately neighbouring bins. A contiguous set of bins with counts greater than 1 is designated a breakpoint region. This results in a set of common breakpoint regions. Each cell is then scored for the occurrence of each of these events, a one meaning the cell has a copy number transition of that type (low to high or high to low) in that genomic region and a zero meaning no copy number transition of that type in that region.

Hierarchical tree of chromosome breakpoints. We used chromosome breakpoints patterns to build a neighbour-joining tree. To eliminate breakpoint events with a high standard deviation, we limited our analysis to breakpoint regions covering no more than seven adjacent bins ($N = 657$). Using a Euclidean metric, we calculated a distance matrix from the binary chromosome breakpoint patterns identified in the single cells using Matlab (Mathworks). From this distance matrix we constructed a tree using average linkage.

Heatmap of chromosome breakpoints. The biclustering heatmap is based on the same set of breakpoints used to build the neighbour-joining tree. Colour indicates the presence of an event, and white means no event. The columns are ordered as in the tree. The rows are events ordered to show clearly which of the subsets of the four main groups share which events. The groups are ordered by subpopulation. A four-dimensional binary vector represents each of the 16 possible subsets of these groups (subset vector). Each breakpoint is represented by a four-dimensional vector of the per cent of cells in each group having an event at that breakpoint (the 'breakpoint vector'). The angle from each breakpoint vector to each subset vector is computed as well as the length of each projection vector. If the length of the projection vector is less than 0.05 the breakpoint vector is assigned to the empty (0,0,0,0) subset, otherwise it is assigned to the subset vector with the smallest angle to the breakpoint vector. The rows are ordered by subset vector in the following order: (1,1,1,1), (0,0,0,1), (0,0,1,0), (0,1,0,0), (1,0,0,0), (0,0,1,1), (0,1,0,1), (1,0,0,1), (0,1,1,0), (1,0,1,0), (1,1,0,0), (0,1,1,1), (1,0,1,1), (1,1,0,1), (1,1,1,0), (0,0,0,0). Within each subset the rows are in descending order by the number of cells in that subset having that event and then in ascending order by the number of cells outside of that subset that do not have that same event.

Analysis of loss of heterozygosity using sequence mutations. PCR duplicates were removed from mapped sequence reads and bases with a quality score below 30 were excluded from analysis. We then determined the set of observed nucleotide types for each cell sequenced from the T10 and T16P and T16M tumours and every position in the genome. For each subpopulation we classified a position as the observed nucleotides only if one or two nucleotide types were each observed in five or more cells in the subpopulation. For each grouping of subpopulations DH, DA, if a classification was made in every subpopulation in the group, we translated the classifications into the generic nucleotides (a,b) based upon the order in which they were seen in the group, from left to right. We counted the resulting classifications of positions for each group by class, and determined whether long blocks of identical classifications along a chromosome were expected by chance. To establish the significance of our classification counts, we repeated our analysis 100 times with randomly permuted cell labels within each group of subpopulations. We eliminated any effects from differing subpopulation size in a separate set of runs of the same analysis, each with 24 randomly selected cells in every subpopulation.

©2011 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature09822

Molecular regulation of sexual preference revealed by genetic studies of 5-HT in the brains of male mice

Yan Liu[1,2]*, Yun'ai Jiang[1,3]*, Yunxia Si[1], Ji-Young Kim[4], Zhou-Feng Chen[4] & Yi Rao[1,5]

Although the question of to whom a male directs his mating attempts[1,2] is a critical one in social interactions, little is known about the molecular and cellular mechanisms controlling mammalian sexual preference. Here we report that the neurotransmitter 5-hydroxytryptamine (5-HT) is required for male sexual preference. Wild-type male mice preferred females over males, but males lacking central serotonergic neurons lost sexual preference although they were not generally defective in olfaction or in pheromone sensing. A role for 5-HT was demonstrated by the phenotype of mice lacking tryptophan hydroxylase 2 (Tph2), which is required for the first step of 5-HT synthesis in the brain. Thirty-five minutes after the injection of the intermediate 5-hydroxytryptophan (5-HTP), which circumvented Tph2 to restore 5-HT to the wild-type level, adult *Tph2* knockout mice also preferred females over males. These results indicate that 5-HT and serotonergic neurons in the adult brain regulate mammalian sexual preference.

Interactions between members of the opposite sex are essential for sexually reproducing animals. Evolutionary benefits have been proposed for homo- and bisexual traits[1,2], which exist in many animals[2] from American bulls[3] to Japanese rhesus monkeys[4]. Studies of animals with different sexual preferences are essential for understanding the seemingly simple decision of a male to court a female.

Research in *Drosophila* has uncovered genes required for *Drosophila* courtship preference, but none of their homologues have been shown to affect mammalian sexual preference. Research in mammals has demonstrated that pheromone sensing in the periphery is important for sexual preference. Male mice lacking *Trpc2* (*Trpc2*$^{-/-}$), which encodes a channel expressed in the vomeronasal organ, mounted other males, emitted ultrasonic vocalizations (USVs) towards males and were less aggressive towards males[5,6]. However, understanding of the central mechanisms for sexual preference remains limited.

The neurotransmitter 5-HT has been implicated in male sexual behaviours such as erection, ejaculation and orgasm in mice and humans[7,8]. Depletion of 5-HT by treating animals with p-chlorophenylalanine (pCPA) or tryptophan-free diets induced male–male mounting[9–11]. However, pCPA treatment was thought to increase sexual activity whereas its effect on sexual preference has not been investigated. Interpretation of pCPA results was complicated further by the lack of specificity: pCPA may affect noradrenaline and dopamine at higher concentrations[12].

Almost all serotonergic neurons in the brain were missing from embryogenesis to adulthood in *Lmx1b* conditional knockout mice in which the floxed *Lmx1b* allele was deleted by *ePet1*-Cre[13]. We compared the behaviours of male mice of different genotypes: *ePet1*-Cre/*Lmx1b*^{flox}/*Lmx1b*^{flox} as homozygous mutants (*Lmx1b*$^{-/-}$); their littermates *ePet1*-Cre/*Lmx1b*$^{flox/+}$ as heterozygous mutants (*Lmx1b*$^{+/-}$); and *Lmx1b*^{flox}/*Lmx1b*^{flox} without *ePet1*-Cre as the wild type (*Lmx1b*$^{+/+}$). We also used *ePet1*-Cre without *Lmx1b*^{flox} as a control.

We tested first how a male responded in his home cage when a wild-type target C57 male was introduced. Compared to the *ePet1*-Cre, *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ controls, *Lmx1b*$^{-/-}$ mice showed significantly more mounting of male intruders (Fig. 1 and Supplementary Movie 1; see Supplementary Data 1 for numbers of mice used and statistics for all figures). The percentage of males who mounted target males was significantly higher in *Lmx1b*$^{-/-}$ males than *ePet1*-Cre, *Lmx1b*$^{+/-}$ and *Lmx1b*$^{+/+}$ males (Fig. 1a). *Lmx1b*$^{-/-}$ males mounted with a shorter latency (Fig. 1b), higher frequency (Fig. 1c) and longer duration (Fig. 1d). These results show that the absence of serotonergic neurons in the brain increased male–male mounting.

A sexually dimorphic behavioural response of males is to emit 30–110 kHz USVs when they encounter female mice or pheromones, which may function as love songs to facilitate female receptivity[14]. *Lmx1b*$^{+/+}$, *Lmx1b*$^{+/-}$ and *Lmx1b*$^{-/-}$ males were similar in USV emission towards females (Fig. 1e–g). However, the percentage of *Lmx1b*$^{-/-}$ males emitting USV towards males was significantly higher than that of *ePet1*-Cre, *Lmx1b*$^{+/+}$ or *Lmx1b*$^{+/-}$ males (Fig. 1f). Numbers of USV 'syllables' emitted towards females were similar among *ePet1*-Cre, *Lmx1b*$^{+/+}$, *Lmx1b*$^{+/-}$ and *Lmx1b*$^{-/-}$ males (Fig. 1g). *Lmx1b*$^{-/-}$ males emitted more USV 'syllables' towards males than *ePet1*-Cre, *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$. The number of USV emissions by *Lmx1b*$^{-/-}$ males towards males was approximately 720 times higher than that of *Lmx1b*$^{+/+}$ males (Fig. 1g).

Although *Lmx1b*$^{-/-}$ males still emitted more USVs towards females, the preference for females over males was significantly reduced: the ratio of USVs towards females over that for males was only 3 for *Lmx1b*$^{-/-}$ males, significantly reduced from 1,002 for *ePet1*-Cre males, 2,438 for *Lmx1b*$^{+/+}$ males and 52 for *Lmx1b*$^{+/-}$.

In the mating choice assay, an oestrous female C57 target mouse and a sexually naive male C57 target mouse were introduced into the home cage of a test male. Wild-type males preferred to mount female targets (Fig. 2a): a higher percentage of *Lmx1b*$^{+/+}$ (or *ePet1*-Cre, *Lmx1b*$^{+/-}$) males mounted female targets than male targets (Supplementary Movie 2). However, the percentage of *Lmx1b*$^{-/-}$ males mounting females was not significantly different from that mounting males. *ePet1*-Cre, *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ males mounted female targets with a shorter latency, higher frequency and longer duration than male targets (Fig. 2b, d, e), whereas *Lmx1b*$^{-/-}$ males mounted males and females with similar latencies, frequencies and durations (Supplementary Movies 2 and 3). Thus, elimination of serotonergic neurons led to a loss of sexual preference in mounting.

Further analyses were carried out to detect a change in sexual preference separate from an increase in sexual drive: (1) in the mating choice assay, all *ePet1*-Cre, *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ males mounted females before males, whereas 46.2% of *Lmx1b*$^{-/-}$ mounted males first (Fig. 2c); (2) the mounting frequency ratio of *Lmx1b*$^{-/-}$ males in the mating choice assay (female mounting frequency − male mounting frequency)/(female + male mounting) (that is, (♀ − ♂/♂ + ♀)) was significantly different from *ePet1*-Cre, *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ males (Fig. 2f); and (3) when a test male was presented only with an oestrous female target, *Lmx1b*$^{-/-}$ males were not statistically significant

[1] National Institute of Biological Sciences, Beijing 102206, China. [2] Graduate School of the Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China. [3] Institute of Neuroscience, Shanghai Institute of Biological Sciences, and Graduate School of the China Academy of Science, China. [4] Departments of Anesthesiology, Psychiatry and Developmental Biology, and the Pain Center, Washington University, School of Medicine, St Louis, Missouri 63110, USA. [5] Peking University School of Life Sciences, State Key Laboratory of Membrane Biology, Beijing 100871, China.
*These authors contributed equally to this work.

©2011 Macmillan Publishers Limited. All rights reserved

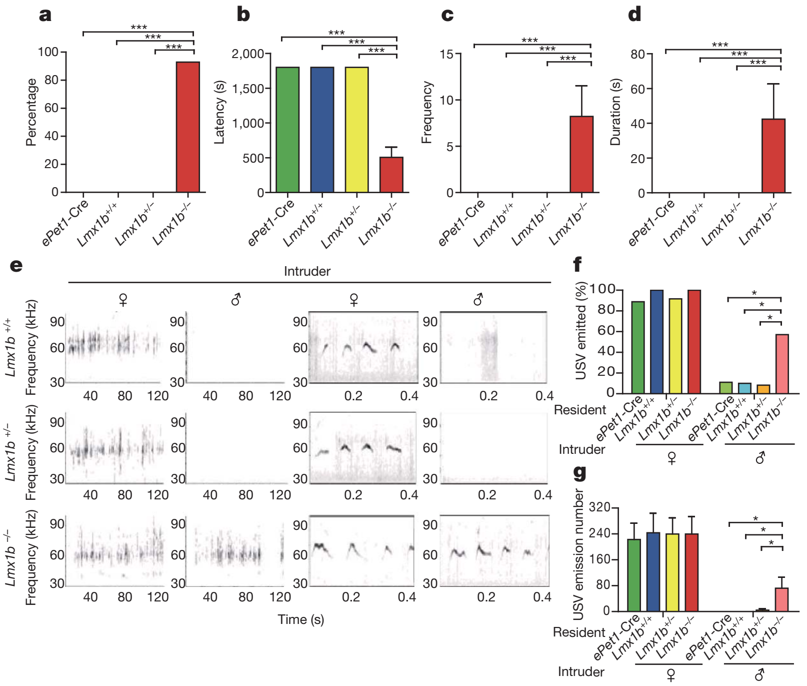


Figure 1 | Male–male mounting and USV by mice lacking central serotonergic neurons. **a**–**g**, Numbers of mice used and statistical analysis are all included in Supplementary Data 1. **a**–**d**, A test male was presented in its home cage with an adult wild-type male and its behaviour was recorded for 30 min (all data shown as mean ± s.e.m.). Compared with *Lmx1b*$^{+/+}$, *Lmx1b*$^{+/-}$ or *ePet1*-Cre, *Lmx1b*$^{-/-}$ males mounted males at a higher percentage (**a**), lower latency (**b**), higher frequency (**c**) and for a longer duration (**d**). **e**, Typical USV patterns emitted by males when presented with female or male intruders. The two left panels show USVs in 2 min, whereas the two right panels show parts of USV graphs at higher magnifications. **f**, Female intruders elicited USV from almost all *ePet1*-Cre, *Lmx1b*$^{-/-}$, *Lmx1b*$^{+/+}$, or *Lmx1b*$^{+/-}$ males . Male intruders elicited USVs more from *Lmx1b*$^{-/-}$ males than from *ePet1*-Cre, *Lmx1b*$^{+/+}$ or *Lmx1b*$^{+/-}$ males. **g**, The number of USVs emitted by *Lmx1b*$^{-/-}$ males towards males is higher than those by *ePet1*-Cre, *Lmx1b*$^{+/+}$ or *Lmx1b*$^{+/-}$ males, whereas *ePet1*-Cre, *Lmx1b*$^{+/+}$, *Lmx1b*$^{+/-}$ and *Lmx1b*$^{-/-}$ males were similar in USVs towards females. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

different from wild-type and heterozygous males in male–female mounting (Supplementary Fig. 1).

We tested male mice for their preference of pheromones present in the genitals or the bedding. In the genital odour preference assay[15], a slide with one half smeared with female genitals and the other half with male genitals was presented to a test male. The total time spent sniffing both halves of the slide was reduced in *Lmx1b*$^{-/-}$ males (Supplementary Fig. 2a). *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ littermates spent significantly more time sniffing female than male genital odour, whereas *Lmx1b*$^{-/-}$ males spent equal time sniffing female and male genital odours (Fig. 3a). *Lmx1b*$^{+/+}$, *Lmx1b*$^{+/-}$ and *Lmx1b*$^{-/-}$ were similar in the amount of time spent sniffing male genital odour. Female genital odour sniffing time was less in *Lmx1b*$^{-/-}$ males than in *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ littermates (Fig. 3a). The genital odour preference ratio (♀ − ♂/♂ + ♀) of *Lmx1b*$^{-/-}$ males was significantly lower than those of *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ males (Fig. 3b). Compared with *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ males, a significantly higher percentage (62.5%) of *Lmx1b*$^{-/-}$ males spent more time sniffing male than female genital odour (Fig. 3c).

In the bedding preference assay[16], the total time spent over male and female bedding was similar among *ePet1*-Cre, *Lmx1b*$^{+/+}$, *Lmx1b*$^{+/-}$ and *Lmx1b*$^{-/-}$ males (Supplementary Fig. 2b). *ePet1*-Cre, *Lmx1b*$^{+/+}$ *and Lmx1b*$^{+/-}$ males spent significantly more time above female than male bedding whereas *Lmx1b*$^{-/-}$ males spent equal time above female and male beddings (Fig. 3d). Compared with *ePet1*-Cre, *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ males, *Lmx1b*$^{-/-}$ males spent more time above male bedding and less time above female bedding. The bedding preference ratio of *Lmx1b*$^{-/-}$ males was significantly lower than those of *ePet1*-Cre, *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ males (Fig. 3e). The percentage of males who spent more time above male bedding was significantly higher in *Lmx1b*$^{-/-}$ males (58.8%) than those in *ePet1*-Cre (0%), *Lmx1b*$^{+/+}$ (6.3%) or *Lmx1b*$^{+/-}$ (12.5%) males (Fig. 3f).

Thus, in both the genital odour and bedding assays, *Lmx1b*$^{-/-}$ males had lost preference for female pheromones over male pheromones: in the genital odour preference assay, *Lmx1b*$^{-/-}$ males showed decreased sniffing time for female genital odour; in the bedding preference assay, *Lmx1b*$^{-/-}$ males showed increased time spent over male bedding and decreased time over female bedding.

Multiple assays involving odour or pheromone sensing were carried out to test for possible changes in olfaction. In the sesame oil preference assay[17], *Lmx1b*$^{+/+}$ and *Lmx1b*$^{-/-}$ males were indistinguishable in spending significantly more time with sesame than air (Supplementary Fig. 3a). In the fox urine avoidance assay[18], *Lmx1b*$^{+/+}$ and *Lmx1b*$^{-/-}$ males were also similar (Supplementary Fig. 3b). Thus, *Lmx1b*$^{-/-}$ males were not defective in either innate attractive or avoidance response.

In the social approach assay[19], *Lmx1b*$^{+/+}$ and *Lmx1b*$^{-/-}$ males were similar in spending more time close to a strange male than the empty chamber (Supplementary Fig. 3c).

In the social recognition assay[20], *Lmx1b*$^{+/+}$ and *Lmx1b*$^{-/-}$ males spent a similar amount of time exploring the first intruder at initial presentation, displayed social habituation towards the familiar intruder over the next three presentations and displayed dishabituation when a new intruder was introduced (Fig. 4a).

An operant conditioning assay was used to test whether *Lmx1b*$^{-/-}$ males could distinguish between male and female pheromones[21]. Two arms of a T maze were supplied with the odour of either female or male urine. Electroshock was applied in such a way that the test mice had to run or stay in the same arm depending on the urine. Over 3 days of training, *Lmx1b*$^{+/+}$ and *Lmx1b*$^{-/-}$ males were similar in learning to avoid punishment (Fig. 4b). Thus, no olfactory defects for general odours or pheromones were detected in *Lmx1b*$^{-/-}$ males.

Results from *Lmx1b*$^{-/-}$ mice indicate a role for serotonergic neurons. To study the role of 5-HT, we used mice unable to synthesize

©2011 Macmillan Publishers Limited. All rights reserved

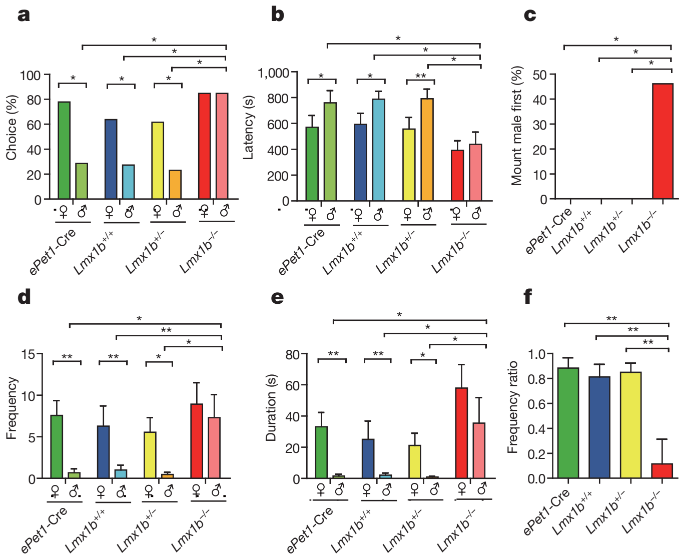


Figure 2 | Lack of sexual preference by mice without central serotonergic neurons. **a**–**f**, Each test male was presented with a male and an oestrous female, and its mating choice was analysed for 15 min. **a**, More *ePet1*-Cre, *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ males mounted female than male targets. A similar percentage of *Lmx1b*$^{-/-}$ males mounted females and males. **b**, *ePet1*-Cre, *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ males mounted female targets faster than male targets. Mounting latencies of *Lmx1b*$^{-/-}$ males for females and males were similar. **c**, More than 40% of *Lmx1b*$^{-/-}$ males but none of the *ePet1*-Cre, *Lmx1b*$^{+/+}$ or *Lmx1b*$^{+/-}$ males chose a male as their first mounting target. **d**, *ePet1*-Cre males mounted females significantly more often than males, as did *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ males. *Lmx1b*$^{-/-}$ males mounted females as often as males ($P > 0.05$, *t*-test). **e**, *ePet1*-Cre males spent more time mounting females than males, as did *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ males. *Lmx1b*$^{-/-}$ males did not show differences in mounting males or females. **f**, The mounting frequency ratio of *Lmx1b*$^{-/-}$ was different from that of *ePet1*-Cre, *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

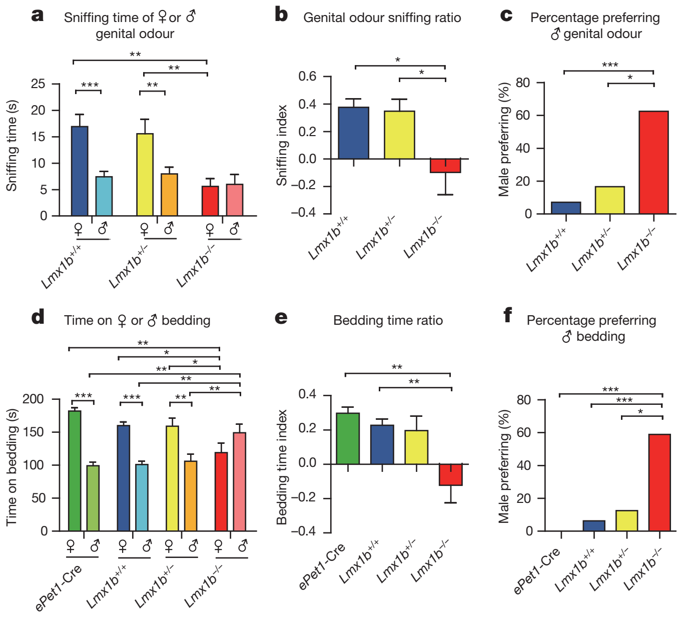


Figure 3 | Loss of sexual preference for genital odour and bedding by males without central serotonergic neurons. **a**, *Lmx1b*$^{+/+}$ males spent more time sniffing female than male genital odour, as did *Lmx1b*$^{+/-}$ males. *Lmx1b*$^{-/-}$ males spent a similar amount of time on female and male genital odour. Three groups were not significantly different in male genital odour sniffing time but *Lmx1b*$^{-/-}$ males spent less time sniffing female genital odour than the other two groups. **b**, Sniffing ratio of *Lmx1b*$^{-/-}$ males was significantly different from *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ males ($P < 0.05$ for *Lmx1b*$^{+/+}$ versus *Lmx1b*$^{-/-}$, $P < 0.05$ for *Lmx1b*$^{+/-}$ versus *Lmx1b*$^{-/-}$, $P > 0.05$ for *Lmx1b*$^{+/+}$ versus *Lmx1b*$^{+/-}$; one-way ANOVA). **c**, Compared with *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$, a higher percentage of *Lmx1b*$^{-/-}$ males spent more time sniffing male than female genital odour. **d**, *ePet1*-Cre males spent more time above female bedding than male bedding, as did *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$ males. *Lmx1b*$^{-/-}$ males spent a similar amount of time above female and male bedding. Compared with *ePet1*-Cre, *Lmx1b*$^{+/-}$ and *Lmx1b*$^{+/+}$, *Lmx1b*$^{-/-}$ males spent less time above female bedding but more time above male bedding. **e**, The bedding time ratio of *Lmx1b*$^{-/-}$ was different from *ePet1*-Cre and *Lmx1b*$^{+/+}$. **f**, Compared with *ePet1*-Cre, *Lmx1b*$^{+/+}$ and *Lmx1b*$^{+/-}$, a significantly higher percentage of *Lmx1b*$^{-/-}$ males spent more time above male bedding. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

5-HT in the brain. 5-HT is synthesized in two steps: tryptophan is converted by a Tph into 5-HTP, which is converted into 5-HT by 5-hydroxytryptophan decarboxylase and aromatic L-amino-acid decarboxylase.

There are two Tph enzymes: Tph2 is required centrally and Tph1 peripherally. We have generated *Tph2*$^{-/-}$ mice (J.-Y.K. *et al*., manuscript in preparation), which were viable[22–24]. High-performance liquid chromatography (HPLC) analysis showed that the 5-HT level was significantly reduced in the brains of *Tph2*$^{-/-}$ males (Supplementary Fig. 4a). Male–male mounting (Supplementary Movie 4) was significantly higher in *Tph2*$^{-/-}$ males than either *Tph2*$^{+/+}$ or heterozygous *Tph2*$^{+/-}$ males: the percentage was significantly higher, duration longer, latency shorter and frequency higher (Supplementary Fig. 4b, c and Fig. 5a, b). In the bedding preference assay, both *Tph2*$^{+/+}$ and *Tph2*$^{+/-}$ males preferred female over male bedding, whereas *Tph2*$^{-/-}$ males showed no preference (Fig. 5c). In the genital odour preference assay, both *Tph2*$^{+/+}$ and *Tph2*$^{+/-}$ males preferred female over male genital odour, but *Tph2*$^{-/-}$ males showed no preference (Fig. 5d).

When presented with an oestrous female target, male–female mounting was not significantly changed in *Tph2*$^{-/-}$ males (Supplementary Fig. 5). In mating choice, *Tph2*$^{-/-}$ males had lost preference for females over males in percentage, latency, frequency and duration (Supplementary Fig. 6a, b, d, e). No control males mounted target males before females, whereas more than 40% of *Tph2*$^{-/-}$ males mounted males first (Supplementary Fig. 6c). The mounting frequency ratio of *Tph2*$^{-/-}$ males was significantly different from those of *Tph2*$^{+/+}$ and *Tph2*$^{+/-}$ males (Supplementary Fig. 6f).

Lmx1b$^{-/-}$ and *Tph2*$^{-/-}$ mice lack 5-HT from embryogenesis. To study the role of 5-HT in adulthood, we took two complementary approaches: first, we depleted 5-HT from adult mice pharmacologically with pCPA[25]; then we attempted to rescue the phenotype of adult *Tph2*$^{-/-}$ mutants.

Adult C57BL/6J males were injected with either pCPA or saline for three consecutive days. 5-HT level was significantly reduced by pCPA

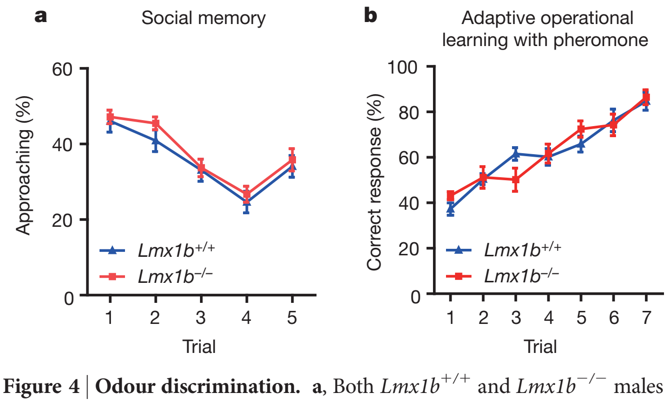


Figure 4 | Odour discrimination. **a**, Both *Lmx1b*$^{+/+}$ and *Lmx1b*$^{-/-}$ males showed habituation and dishabituation in sniffing time. No statistical difference was found between *Lmx1b*$^{+/+}$ and *Lmx1b*$^{-/-}$ males at any point. **b**, After seven training sessions with male and female urine, no significant difference was detected between *Lmx1b*$^{+/+}$ and *Lmx1b*$^{-/-}$ males at any point.

©2011 Macmillan Publishers Limited. All rights reserved

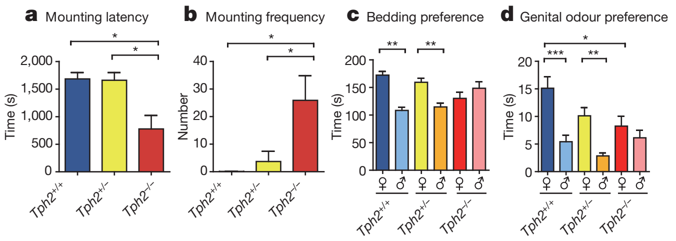


Figure 5 | Brain chemistry and behaviours of *Tph2* knockout males. **a**, **b**, Compared with *Tph2*$^{+/+}$ and *Tph2*$^{+/-}$, *Tph2*$^{-/-}$ males showed a shorter latency (**a**) and higher frequency in mounting males (**b**). **c**, Both *Tph2*$^{+/+}$ and *Tph2*$^{+/-}$ males significantly preferred female over male bedding, whereas *Tph2*$^{-/-}$ males did not show a preference between male and female bedding. **d**, Both *Tph2*$^{+/+}$ and *Tph2*$^{+/-}$ males significantly preferred female over male genital odour, whereas *Tph2*$^{-/-}$ males did not show a preference between male and female genital odour. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

(Supplementary Fig. 7). pCPA-treated males showed shorter latency, higher frequency and longer duration than control males in mounting target males (Supplementary Fig. 8a–d), and lost bedding preference (Supplementary Fig. 8e, f).

To test whether 5-HTP injection into adult mice could rescue the *Tph2*$^{-/-}$ phenotype, we examined first whether 5-HTP could rescue 5-HT synthesis in *Tph2*$^{-/-}$ males and found that 5-HT levels were restored 35 min after intraperitoneal injection of 5-HTP but not saline (Fig. 6a and Supplementary 9a, b).

5-HTP significantly reduced male–male mounting of *Tph2*$^{-/-}$ males: the percentage was decreased, latency increased, frequency decreased and duration shortened; all returning to wild-type levels (Fig. 6b, c and Supplementary Fig. 9c, d). 5-HTP rescued the loss of sexual preference in mounting latency, frequency and duration in the mating choice assay (Supplementary Fig. 10a–c) and the bedding preference of *Tph2*$^{-/-}$ males (Fig. 6d and Supplementary Fig. 9e).

When a test male was presented with a target female, *Tph2*$^{-/-}$ males were similar to wild-type and heterozygous males in mounting percentage, latency, frequency and duration (Supplementary Figs 5, 11). 5-HTP injection into *Tph2*$^{-/-}$ males did not affect male–female mounting (Supplementary Fig. 11), although 5-HTP injection into wild-type males reduced male–female mounting. Because 5-HTP injection in wild-type males increased the level of 5-HT beyond the wild-type level (Supplementary Fig. 9a, b), it indicated a dosage-sensitive effect of 5-HT: 5-HT at concentrations above the wild-type level inhibited male–female mounting, but 5-HT concentrations between the wild-type and *Tph2*$^{-/-}$ levels did not affect male–female mounting.

We conclude that central serotonergic signalling is crucial for male sexual preference in mice. This is the first time, to our knowledge, that a neurotransmitter in the brain has been demonstrated to be important in mammalian sexual preference. Previous studies in mammals have implicated 5-HT and dopamine in male sexual behaviours, but neither has been demonstrated to have any role in sexual preference: dopamine is thought to facilitate male sexual behaviours whereas 5-HT is thought to inhibit sexual behaviours[7–11,26]. Our studies have established a role for 5-HT in male sexual preference. Multiple results showed a loss in sexual preference beyond or separate from hypersexuality: (1) the ratio of male–male and male–female interactions was repeatedly measured to analyse sexual preference (Figs 2f, 3b, e, 5c, d, 6d and Supplementary Figs 6f, 8f, 9e, 10d); (2) *Lmx1b*$^{-/-}$ males showed increased USVs towards males but not towards females (Fig. 1g); (3) in mating choice, the latency, frequency and duration of *Lmx1b*$^{-/-}$ males to mount males, but not to mount females, was changed (Fig. 2a, b, d, e); (4) in bedding preference, *Lmx1b*$^{-/-}$ (Fig. 3d) and *Tph2*$^{-/-}$ males (Figs 5c, 6d) showed an increase in time spent over male bedding but a decrease in time over female bedding; (5) wild-type males always mounted females before males but a significant fraction of *Lmx1b*$^{-/-}$ or *Tph2*$^{-/-}$ males mounted males first (Fig. 2c and Supplementary Fig. 6c); (6) in the genital odour preference assay, both *Lmx1b*$^{-/-}$ (Fig. 3a) and *Tph2*$^{-/-}$ (Supplementary Fig. 5d) males showed a decrease in time on female genital odour, which could not be explained by hypersexuality; and (7) when presented with an oestrous target female, neither *Lmx1b*$^{-/-}$ males (Supplementary Fig. 1) nor *Tph2*$^{-/-}$ males (Supplementary Fig. 5) were different from wild-type males.

Increased sexual drive was observed in males lacking 5-HT when they were tested in the presence of live target males and females (Supplementary Fig. 6). This has been noted before in mice defective for Trpc2 and vomeronasal organ olfaction[5,6]. *Trpc2*$^{-/-}$ males have been previously reported to have lost male–female preference in mating choice[5,6]. *Trpc2*$^{-/-}$ males showed increased mounting of both males and females (figure 2c in ref. 6). The conclusion of a loss in sexual preference in *Trpc2*$^{-/-}$ males was inferred from a relative change: *Trpc2*$^{-/-}$ males showed a 2-fold preference for females over males whereas the wild-type showed a 10-fold preference. The phenotypes reported here for *Lmx1b*$^{-/-}$, *Tph2*$^{-/-}$ males and pCPA-treated males were stronger than for *Trpc2*$^{-/-}$ males in mating choice: these males did not show significant preference for females (Fig. 2 and Supplementary Fig. 6).

At present, it is not known whether 5-HT regulates the vomeronasal organ pathway in pheromone sensing or acts further downstream in behavioural decisions. Differences have been noted between *Trpc2* and *Lmx1b* in the brain: aggression was largely lost in *Trpc2*$^{-/-}$, but not *Lmx1b*$^{-/-}$, mice (data not shown). It is more likely that 5-HT regulates central decision-making than influencing peripheral olfaction. However, we cannot completely rule out the possibility that 5-HT regulates a specific innate olfactory pathway processing sexual information[27]. In mice, it will be interesting to identify specific subsets of serotonergic neurons and serotonergic receptors involved in sexual preference.

An unavoidable question raised by our findings is whether 5-HT has a role in sexual preference in other animals. In a positron emission tomography study of humans, the response of heterosexual men to the selective serotonin reuptake inhibitor (SSRI) fluoxetine was found to be different from that of homosexual men[28]. SSRIs inhibited compulsive sexual behaviours in homosexual and bisexual men[29]. However, so far, none of these studies has investigated whether 5-HT has a role in

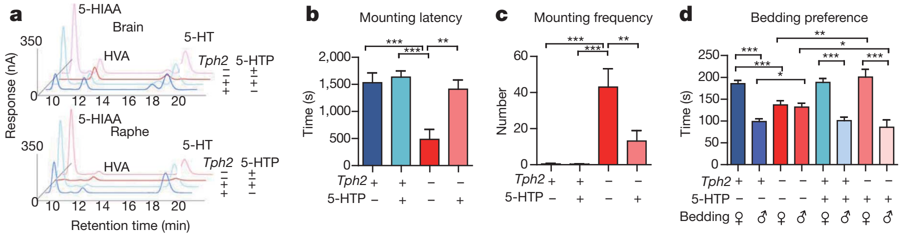


Figure 6 | 5-HTP rescue of chemical and behavioural deficits in *Tph2* knockout mice. **a**, Levels of 5-HT and 5-hydroxyindoleacetic acid (5-HIAA) were analysed in *Tph2*$^{+/+}$ and *Tph2*$^{-/-}$ males 35 min after injection of either 5-HTP (40 mg kg^{-1} body weight) or control saline. **b**, **c**, Male–male mounting in *Tph2*$^{-/-}$ mice was significantly rescued by 5-HTP: the latency was lengthened and frequency reduced. **d**, Bedding preference was monitored between 35 and 40 min after injection. 5-HTP could significantly restore the preference of female over male bedding by *Tph2*$^{-/-}$ males.

©2011 Macmillan Publishers Limited. All rights reserved

sexual preference. Attempts have been made to map genetic loci affecting human sexuality[30], although specific genes have not been identified. Our discovery of a role for serotonergic signalling in mouse sexual preference should stimulate further studies into the role of 5-HT in sexual interactions in particular and roles of neurotransmitters in mammalian social relationships in general.

METHODS SUMMARY

We used conditional knockout mice for *Lmx1b* and knockout mice for *Tph2*. Levels of 5-HT in these mice and their heterozygous and wild-type littermates were measured by HPLC. Most of the behavioural assays were similar to established methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 12 August 2010; accepted 14 January 2011.

Published online 23 March; corrected 7 April 2011 (see full-text HTML version for details).

1. Trivers, R. L. Parent–offspring conflict. *Am. Zool.* **14,** 249–264 (1974).
2. Sommer, V. & Vasey, P. L. *Homosexual Behaviour in Animals: An Evolutionary Perspective* (Cambridge Univ. Press, 2006).
3. Price, E. O. & Wallach, S. J. Development of sexual and aggressive behaviors in Hereford bulls. *J. Anim. Sci.* **69,** 1019–1027 (1991).
4. Erwin, J. & Maple, T. Ambisexual behavior with male–male anal penetration in male rhesus monkeys. *Arch. Sex. Behav.* **5,** 9–14 (1976).
5. Stowers, L., Holy, T. E., Meister, M., Dulac, C. & Koentges, G. Loss of sex discrimination and male–male aggression in mice deficient for TRP2. *Science* **295,** 1493–1500 (2002).
6. Leypold, B. G. *et al.* Altered sexual and social behaviors in trp2 mutant mice. *Proc. Natl Acad. Sci. USA* **99,** 6376–6381 (2002).
7. Hull, E. M., Muschamp, J. W. & Sato, S. Dopamine and serotonin: influences on male sexual behavior. *Physiol. Behav.* **83,** 291–307 (2004).
8. Hull, E. M. & Dominguez, J. M. Sexual behavior in male rodents. *Horm. Behav.* **52,** 45–55 (2007).
9. Ferguson, J. *et al.* ''Hypersexuality'' and behavioral changes in cats caused by administration of p-chlorophenylalanine. *Science* **168,** 499–501 (1970).
10. Malmnäs, C. & Meyerson, B. p-Chlorophenylalanine and copulatory behaviour in the male rat. *Nature* **232,** 398–400 (1971).
11. Salis, P. & Dewsbury, D. p-Chlorophenylalanine facilitates copulatory behaviour in male rats. *Nature* **232,** 400–401 (1971).
12. Dailly, E., Chenu, F., Petit-Demouliere, B. & Bourin, M. Specificity and efficacy of noradrenaline, serotonin depletion in discrete brain areas of Swiss mice by neurotoxins. *J. Neurosci. Methods* **150,** 111–115 (2006).
13. Zhao, Z.-Q. *et al. Lmx1b* is required for maintenance of central serotonergic neurons and mice lacking central serotonergic system exhibit normal locomotor activity. *J. Neurosci.* **26,** 12781–12788 (2006).
14. Guo, Z. & Holy, T. E. Sex selectivity of mouse ultrasonic songs. *Chem. Senses* **32,** 463–473 (2007).
15. Ferkin, M. H. & Li, H. Z. A battery of olfactory-based screens for phenotyping the social and sexual behaviors of mice. *Physiol. Behav.* **85,** 489–499 (2005).
16. Moncho-Bogani, J., Lanuza, E., Herndez, A., Novejarque, A. & Martez-Garc, F. Attractive properties of sexual pheromones in mice: innate or learned? *Physiol. Behav.* **77,** 167–176 (2002).
17. Burwash, M. D., Tobin, M. E., Woolhouse, A. D. & Sullivan, T. P. Laboratory evaluation of predator odors for eliciting an avoidance response in roof rats (*Rattus rattus*). *J. Chem. Ecol.* **24,** 49–66 (1998).
18. Blanchard, D. *et al.* Failure to produce conditioning with low-dose trimethylthiazoline or cat feces as unconditioned stimuli. *Behav. Neurosci.* **117,** 360–368 (2003).
19. Nadler, J. J. *et al.* Automated apparatus for quantitation of social approach behaviors in mice. *Genes Brain Behav.* **3,** 303–314 (2004).
20. Ferguson, J. N. *et al.* Social amnesia in mice lacking the oxytocin gene. *Nature Genet.* **25,** 284–288 (2000).
21. Yan, Z. *et al.* Precise circuitry links bilaterally symmetric olfactory maps. *Neuron* **58,** 613–624 (2008).
22. Gutknecht, L. *et al.* Deficiency of brain 5-HT synthesis but serotonergic neuron formation in *Tph2* knockout mice. *J. Neural Transm.* **115,** 1127–1132 (2008).
23. Savelieva, K. V. *et al.* Genetic disruption of both tryptophan hydroxylase genes dramatically reduces serotonin and affects behavior in models sensitive to antidepressants. *PLoS ONE* **3,** e3301 (2008).
24. Alenina, N. *et al.* Growth retardation and altered autonomic control in mice lacking brain serotonin. *Proc. Natl Acad. Sci. USA* **106,** 10332–10337 (2009).
25. Koe, B. K. & Weissman, A. p-Chlorophenylalanine: a specific depletor of brain serotonin. *J. Pharmacol. Exp. Ther.* **154,** 499–516 (1966).
26. Gawienowski, A. M. & Hodgen, G. D. Homosexual activity in male rats after p-chlorophenylalanine: effects of hypophysectomy and testosterone. *Physiol. Behav.* **7,** 551–555 (1971).
27. Kobayakawa, K. *et al.* Innate versus learned odour processing in the mouse olfactory bulb. *Nature* **450,** 503–508 (2007).
28. Kinnunen, L., Moltz, H., Metz, J. & Cooper, M. Differential brain activation in exclusively homosexual and heterosexual men produced by the selective serotonin reuptake inhibitor, fluoxetine. *Brain Res.* **1024,** 251–254 (2004).
29. Wainberg, M. *et al.* A double-blind study of citalopram versus placebo in the treatment of compulsive sexual behaviors in gay and bisexual men. *J. Clin. Psychiatry* **67,** 1968–1973 (2006).
30. Mustanski, B. S. *et al.* A genomewide scan of male sexual orientation. *Hum. Genet.* **116,** 272–278 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to E. S. Deneris for *ePet1*-Cre mice; to R. Johnson for *Lmx1b*^{fl} mice; to M. Luo for discussions; to Z. Yan and Y. Lu for the operant conditioning apparatus; to X. Wang and Y. Wan for help with HPLC; to J. Lang and J. Yin for mouse breeding and genotyping; to P. Ding, P. Wang, H. Lu and X. Wang for technical assistance; to L. Zhao, Z. Qiu and H. Jing for animal caring; and to the Ministry of Science and Technology (973 program 2010CB833901) and Beijing Municipal Commission on Science and Technology for grant support (to Y.R.), and the NIH for grant support (to Z.-F.C.).

Author Contributions Y.R. conceived the project, Y.R., Y.L. and Y.J. designed the experiments, Y.L., Y.J. and Y.S. performed the experiments, J.-Y.K. and Z.-F.C. contributed the Tph2 knockout mutants, Y.R., Y.L. and Y.J. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to Y.R. (yrao@pku.edu.cn).

©2011 Macmillan Publishers Limited. All rights reserved

METHODS

Mouse stocks. *ePet1*-Cre mice were a gift from E. S. Deneris and the floxed *Lmx1b* mice were a gift from R. Johnson. *Tph2* knockout mice were generated by deleting exon 5, which encodes the tryptophan hydroxylase domain (for details see J.-Y.K. *et al.*, manuscript submitted). Mice were weaned at the age of 21 days. Mice were maintained on a 12 h light, 12 h dark schedule and housed initially in groups of five up to the tenth week and then singly housed until the end of experiments. Food and water were provided *ad libitum*. Room temperature was 23 ± 1 °C. Humidity was 40–60%. All test mice were 12–16 weeks old. The target mice were 11–13 weeks old.

Mouse genotyping. Genomic DNA was extracted from mouse tail tissues at the day of weaning. Mutant mice were generated by crossing *ePet1*-Cre mice with floxed *Lmx1b* mice and following intercross within the F1 generation mice. Littermates used in the tests were of the same sex and similar body weight as the knockout mice. The primers were: AGGCTCCATCCATTCTTCTC (floxed *Lmx1b1*); CCACAATAAGCAAGAGGCAC (floxed *Lmx1b2*); ATTTGCCTGCA TTACCGGTCG (Cre1); CAGCATTGCTGTCACTTGGTC (Cre2).

Immunocytochemical analysis with anti-5-HT antibodies confirmed that 5-HT-positive neurons were absent in *Lmxb1* knockout mice (data not shown).

The *Tph2* line was maintained by crossing heterozygotes. Littermates included wild-type, heterozygotes and homozygous knockout mice. The primers for genotyping were: GGGCATCTCAGGACGTAGTAG; GGGCCTGCCGATAGTAA CAC; GCAGCCAGTAGACGTCTCTTAC.

Measurement of 5-HT. The levels of 5-HT and its metabolites were separated by HPLC and measured by an electrochemical detector in samples from adult male mice. In 5-HTP rescue experiments, mice were injected with 40 mg kg^{-1} 5-HTP or saline (both at the volume of 5 ml kg^{-1}). They were euthanized 35 min later. The brain was dissected and the raphe region was isolated on ice. Samples were weighed before ultrasonication. Monoamines were extracted by perchloric acid. The sample was filtrated by 0.22 μm filter before being injected into RP-HPLC (ESA). Noradrenaline, 3,4-dihydroxyphenylacetic acid (DOPAC), dopamine, HIAA, homovanillic acid (HVA) and 5-HT were measured by an electrochemical detector. Their concentrations were calculated by CoulArray software (ESA) based on standard samples. Values of amine per wet tissue weight are shown in the final figures.

Order of behavioural assays. Male mutant mice and their littermates at 12–13 weeks of age and of similar body weight were sexually naive and group-housed with same-sex mice before 10 weeks of age. After 2 weeks of single housing, mice were tested in the following order: bedding preference, male–male resident–intruder assay, mating choice assay, sexual behaviours with an oestrous female, bedding preference again (no difference was observed with results from the first bedding preference). Mice were given one week of rest between each test. For *Lmx1b* mice, the same group of mice were used in male–male mounting, mating choice and male–female mounting. For *Tph2* mice, a different group were used for male–female mounting. Sexually experienced mice were used for USV, social approach, habituation and olfactory learning assays. Sexually naive mice were used for urine preference and olfactory tests.

Resident–intruder tests. All test mice were sexually naive. The bedding of the test mice had not been changed for at least 4 days. Intruder mice were 11–13 weeks old, sexually naive and group-housed C57Bl/6J males. All activities within a test were recorded by an infrared camera (Sony Video Recorder, DCR-HC26C). Mounting latency, mounting frequency and total duration of mounting within 30 min were measured.

Mating choice assay. Beddings of test mice had not been changed for at least four days. A group-housed sexually naive 11–13 week-old C57Bl/6J male and a sexually naive oestrous 10-week-old female C57Bl/6J female were introduced into the cage of each test male. Each assay lasted 15 min after the target mice were introduced. All activities were recorded by an infrared camera. The latency, frequency and duration of mounting of male or female targets were analysed.

Sexual behaviours with females. An oestrous female was presented to a test male and video was recorded for 30 min using an infrared camera. The latency, frequency and duration of male mounting of the female were analysed.

USVs. Tests were carried out with singly housed adult males during the dark phase in the home cage. UltraSoundGate 116-200 system (Avisoft) was used to record the ultrasound. We recorded the background sound for 1 min before a stimulus mouse of 10–13 weeks old was introduced. The recording lasted for 2 min. Recorded data was analysed with SASLab (Avisoft)[5]. Sounds over the frequency range of 30–110 kHz were analysed. Profiles of background noise created by mouse movement were very different from USVs. To confirm that the resident mouse was the source of USVs, we recorded from assays in which either the resident or the intruder mouse was devocalized. We were able to record robust USVs (presented in our figures) only when the intruder mouse was devocalized and not when the resident mouse was devocalized.

Genital odour preference assay. This assay was modified from a previously described procedure[15]. The anogenital area scent from a male was rubbed on the left or right side of a clean glass microscope slide while the anogenital area scent from a female was rubbed on the other side of the slide. Five seconds later, the slide was hung in the middle of the cage by a clamp. The slides were ~5 cm over the bedding. Activities of the test mice were recorded for 3 min by an infrared camera and the sniff time on the scent portion on either side was analysed as was the amount of time a test male licked the slide or its nose touched the slide.

Bedding preference assay. Bedding from group-housed adult C57BJ/6J males or females was not changed for 4 days. Ten grams of male or female bedding were put in one side on the bottom of a cage in an area of 11.5 × 17 cm^2. Male and female beddings were prevented from mixing by a plastic bar of 6 cm. The size of cage was 29 × 17 × 15 cm (length × width × height)[16]. A grid of plastic bars separated the test mice from the bedding on the bottom of the cage. The bars were 5 mm wide with 5 mm intervals. The test mouse was put into the cage to be familiarized with the cage without bedding for 5 min before the mice were taken out and the bedding and a clean grid was put into the cage. After each assay, the cage was washed with water and then alcohol to remove odour.

Olfactory learning assay. We employed a T maze in which electric shock could be applied to either side of the horizontal chamber as described previously[21]. Briefly, there was a door at the intersection of the horizontal and vertical chambers. The horizontal chamber of 8 × 8 × 60 cm^3 was divided into three parts: a left arm of 8 × 8 × 23 cm, a right arm of 8 × 8 × 23 cm and a middle zone of 8 × 8 × 14 cm. Each test mouse was introduced into the vertical chamber of the T maze. After it entered the horizontal chamber, the door between the vertical and horizontal chambers was closed and the mouse was allowed to walk within the horizontal chamber. The mouse was not allowed to stay in the middle zone for longer than 8 s, otherwise it would be punished with electroshock. The position of the test mouse was monitored by a video recorder. Urine samples were collected from more than 20 C57BL/6J males or females and stored at −20 °C. A 1.5 ml urine sample was used for each test. The odour of male or female urine was puffed into the left or right arm of the horizontal chamber and expirated from the middle zone. Odour was presented for 50 s. We trained the test male mouse with electroshock to stay in the arm with female odour and to avoid the arm with male odour. The mouse had to make a decision to stay in or leave the arm when an odour was presented. Each training session of 18 trials lasted for 30 min. Every mouse was given 6 training sessions over 3 days before the final test. There were 10 trials in the final test. The percentages of correct choices in every training session and the final test were analysed.

Innate behavioural responses to odours. The set-up is the same as that for the olfactory learning assay, except that no electroshock was applied. Sexually naive males (mutants or littermates) of 10–16 weeks old were tested for their choices of fox urine versus air, or sesame oil versus air. Fox urine was used to test the innate avoidance of a predator's odour. Fox urine was diluted at two concentrations (60× and 20×). The main air flow velocity was 250 l h^{-1}. The air flow through fox urine was 70 ml min^{-1} or 210 ml min^{-1}, respectively. The time that mice spent in the empty arm or the fox urine arm was recorded by Matlab software. Sesame oil diluted 83× was used to test innate attraction to food. Time spent in the air arm or the sesame oil arm was recorded by Matlab software.

Social approach. The social approach experiment was tested in a modified T-shaped box. There was a small cage separated by wire at each end of the arms in the horizontal chamber. A test mouse was allowed to habituate for five minutes before an unfamiliar target male was randomly placed in one of the small cages. The target mouse could be seen, smelled and heard, but could not be touched. The test mouse was allowed to move in the box for 5 min. Its location was video recorded and analysed by a computer.

Social memory. Singly housed adult males were tested in the dark phase and in the room where they were reared. Ovariectomized C57Bl/6J females were used as stimulus mice[20]. They were ovariectomized at 6 weeks old and used 2 weeks later. A stimulus mouse was introduced into the cage housing a test mouse for 1 min and then was removed. After an interval of 10 min, the same stimulus female was introduced again for 1 min. The stimulus mouse was presented four times. On the fifth time, a new stimulus mouse was introduced for 1 min. The behaviour of test mice was videotaped and time spent on body sniffing was analysed.

5-HT depletion by pCPA treatment. Male C57Bl/6J mice of 11–13 weeks of age were used. They were injected with either 500 mg kg^{-1} of pCPA (Sigma, C6506) or saline control for 3 consecutive days after 4 days of being singly housed. Animals were tested with adult C57 female mice. Mice that did not show mounting behaviour in 15 min were discarded. Mice that qualified were then singly housed for 1 week before social behaviour testing and their bedding was not changed. Animals were randomly divided into pCPA or saline treatment groups. pCPA was suspended in 1% Tween saline at a concentration of 50 mg ml^{-1}. The pCPA group were injected intraperitonially with pCPA (10 ml kg^{-1}) at 72, 48 and 24 h before testing. The control group received 1% Tween saline. Resident–intruder and mating choice assays were carried out. Behavioural tests were performed in the dark.

©2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09823

Local, persistent activation of Rho GTPases during plasticity of single dendritic spines

Hideji Murakoshi[1], Hong Wang[1] & Ryohei Yasuda[1,2]

The Rho family of GTPases have important roles in the morphogenesis of the dendritic spines[1–3] of neurons in the brain and synaptic plasticity[4–9] by modulating the organization of the actin cytoskeleton[10]. Here we used two-photon fluorescence lifetime imaging microscopy[11–13] to monitor the activity of two Rho GTPases—RhoA and Cdc42—in single dendritic spines undergoing structural plasticity associated with long-term potentiation in CA1 pyramidal neurons in cultured slices of rat hippocampus. When long-term volume increase was induced in a single spine using two-photon glutamate uncaging[14,15], RhoA and Cdc42 were rapidly activated in the stimulated spine. These activities decayed over about five minutes, and were then followed by a phase of persistent activation lasting more than half an hour. Although active RhoA and Cdc42 were similarly mobile, their activity patterns were different. RhoA activation diffused out of the stimulated spine and spread over about 5 μm along the dendrite. In contrast, Cdc42 activation was restricted to the stimulated spine, and exhibited a steep gradient at the spine necks. Inhibition of the Rho–Rock pathway preferentially inhibited the initial spine growth, whereas the inhibition of the Cdc42–Pak pathway blocked the maintenance of sustained structural plasticity. RhoA and Cdc42 activation depended on Ca^{2+}/calmodulin-dependent kinase (CaMKII). Thus, RhoA and Cdc42 relay transient CaMKII activation[13] to synapse-specific, long-term signalling required for spine structural plasticity.

Previous studies using two-photon fluorescence lifetime imaging microscopy (2pFLIM) and two-photon glutamate uncaging revealed the spatiotemporal dynamics of the signalling proteins CaMKII and HRas (also known as transforming protein 21) in single spines undergoing structural plasticity and long-term potentiation[12,13]. CaMKII activation is restricted to spines, and decays rapidly with a time constant of about ten seconds[13]. In contrast, HRas activity spreads from the stimulated spines along dendrites and into surrounding spines over about 10 μm (ref. 12). However, to achieve long-lasting, spine-specific plasticity, there should also exist signalling pathways that relay compartmentalized signalling on the timescale of minutes to hours. Rho GTPases may constitute such signalling, because they are important in regulating the actin cytoskeleton[3,16], which is essential for spine-specific, long-term structural and functional plasticity[14,17].

To measure the activation of Rho GTPases in single dendritic spines, we developed fluorescence resonance energy transfer (FRET)-based sensors optimized for imaging under 2pFLIM using a design similar to a previously developed HRas sensor[11]. The RhoA/Cdc42 sensors consist of two components: RhoA/Cdc42 tagged with monomeric enhanced green fluorescent protein (mEGFP) and their binding partner, Rho GTPase binding domain (RBD) of Rhotekin/Pak3, doubly tagged with mCherry (mCherry–RBD–mCherry) (Supplementary note). When mEGFP–Rho GTPase is activated, mCherry–RBD–mCherry binds to mEGFP–RhoA/Cdc42, causing FRET between mEGFP and mCherry (Supplementary Figs 1 and 2). These sensors were verified to be specific and sensitive under 2pFLIM (Supplementary note).

Using these sensors, we measured the activity of RhoA and Cdc42 during spine structural plasticity associated with long-term potentiation

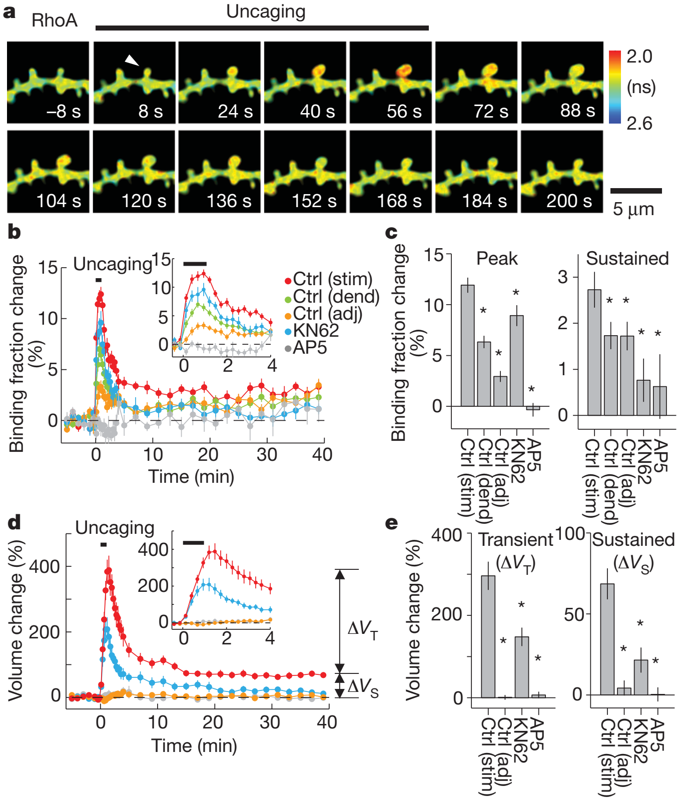


Figure 1 | Spatiotemporal dynamics of RhoA activation during long-term structural plasticity induced in single spines **a**, Fluorescence lifetime images of RhoA activation during spine structural plasticity induced by two-photon glutamate uncaging. Arrowhead indicates the stimulated spine. Warmer colours indicate shorter lifetimes and higher RhoA activity. Scale bar, 5 μm. **b**, Time course of RhoA activation measured as a change in the fraction of mEGFP–RhoA bound to mCherry–RBD–mCherry in stimulated spines (stim), the dendritic shaft beside the stimulated spines (dend; within 1 μm), and adjacent spines (adj; between 3–5 μm of the stimulated spines). Data using pharmacological inhibitors (Ctrl, control condition; KN62, CaMKII inhibitor; AP5, NMDA receptor inhibitor) are also shown. The inset to **b** shows a closer view of the first 4 min. The numbers of samples (spines/neurons) are 35/29 for stimulated spines and dendrites, 29/26 for adjacent spines, 16/10 for KN62 and 8/5 AP5. Error bars are s.e.m. **c**, Transient (averaged over 16–64 s) and sustained (averaged over 20–38 min) RhoA activation. Stars denote statistically significant difference (<0.05) from the value in the stimulated spines under the control condition. Wilcoxon signed-rank test was used for dendrites and adjacent spines, and analysis of variance (ANOVA) followed by post-hoc tests using the least significant difference was used for experiments with pharmacological inhibitors. **d**, Averaged time course of spine volume change in the same experiments as in **b**. The inset to **d** shows a closer view of the first 4 min. **e**, Transient (volume change averaged over 1.5–2 min subtracted by that over 20–38 min) and sustained volume change (volume change averaged over 20–38 min).

[1] Department of Neurobiology, [2] Howard Hughes Medical Institute, Duke University Medical Center, Durham, NC 27710, USA.

©2011 Macmillan Publishers Limited. All rights reserved

(Figs 1, 2 and 3). Pyramidal neurons in the CA1 region of cultured hippocampal slices were ballistically[18] transfected with the RhoA or Cdc42 sensor, and the FRET signal was imaged under 2pFLIM. The spine volume was monitored using the red fluorescence of mCherry–RBD–mCherry (Supplementary Fig. 3)[12]. To induce structural plasticity in a single dendritic spine, we applied a low-frequency train of two-photon glutamate uncaging pulses (30 pulses at 0.5 Hz) to the spine in zero extracellular Mg^{2+} (refs 13, 14 and 19). The spine volume increased rapidly by about 300% following glutamate uncaging (transient phase) and relaxed to an elevated level of 70–80% for more than 30 min (sustained phase) (Figs 1d and 2d)[12–14]. The time course of spine enlargement in neurons expressing the FRET sensor was similar to that in neurons expressing only EGFP (Fig. 4)[14], suggesting that the overexpression of FRET sensors causes almost no effects on spine structural plasticity (Supplementary note).

Under basal conditions, there was no correlation between the activity of Rho GTPases and spine volume (Supplementary Fig. 4). When spine structural plasticity was induced, both RhoA (Fig. 1a and b) and Cdc42 (Fig. 2a and b) were activated rapidly within about 30 s in the stimulated spines. The activation decayed over about 5 min, followed by sustained activity lasting more than 30 min. RhoA activation spread into the dendrites over several micrometres (Figs 1a–c, and 3a and b), and invaded surrounding spines to a small extent (∼25% of the stimulated spines). In contrast, Cdc42 activation was restricted to the stimulated spines (Figs 2a–c, and 3c and d). For both RhoA and Cdc42, the gradient at the spine necks was maintained for more than about 30 min (Figs 1b and c, and 2b and c).

Next, we pharmacologically identified the signalling pathways that activate RhoA and Cdc42. Inhibition of *N*-methyl-D-aspartic acid (NMDA) receptors with 2-amino-5-phosphonopentanoic acid (AP5, 50 μM) abolished activation of RhoA (Fig. 1b and c) and Cdc42 (Fig. 2b and c) as well as spine enlargement (Figs 1d and e, and 2d and e)[14], indicating that RhoA and Cdc42 are activated by Ca^{2+} through NMDA receptors. The CaMKII inhibitor KN62 is known to strongly inhibit sustained spine enlargement, but has significantly less of an effect on transient spine enlargement (Figs 1d and e, and 2d and e)[13,14]. KN62 (10 μM) partially inhibited RhoA and Cdc42 activation during the transient phase, and more strongly during the sustained phase. (Figs 1b and c, and 2b and c). Expression of autocamtide CaMKII inhibitor peptide 2 (AIP2) also inhibited spine volume change and Rho GTPase activation in a similar manner (Supplementary Fig. 5). These results suggest that RhoA and Cdc42 are downstream of CaMKII.

We next characterized the spatial profile of RhoA and Cdc42 activities along dendrites as a function of the distance from the stimulated spines (Fig. 3). RhoA activity showed a relatively small gradient between the stimulated spines and dendrites, and spread along the dendrites. The length constant of spread along the dendrite was 4.5 μm (Fig. 3a and b), a value similar to that for HRas (about 10 μm)[12]. In contrast, Cdc42 activation was restricted to the stimulated spines (Fig. 3c and d), showing a spatial pattern similar to that of CaMKII[13]. A small fraction of Cdc42 activation spread into the

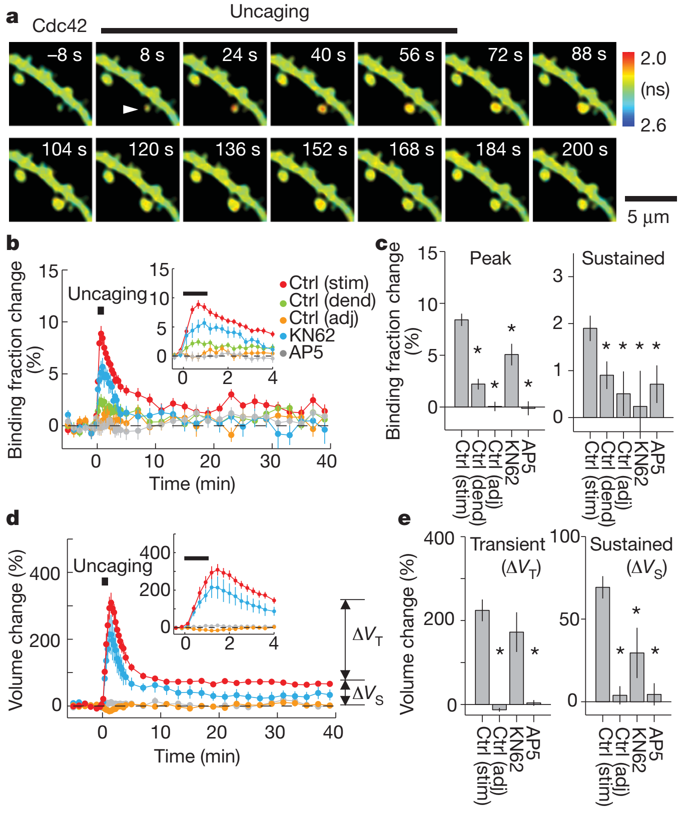


Figure 2 | Spatiotemporal dynamics of Cdc42 activation during long-term structural plasticity induced in single spines. The same experiments and analyses as in Fig. 1 but measuring Cdc42 activity instead of RhoA activity. The numbers of samples (spines/neurons) are 33/28 for stimulated spines and dendrite, 33/28 for adjacent spines, 11/6 for KN62 and 12/8 for AP5.

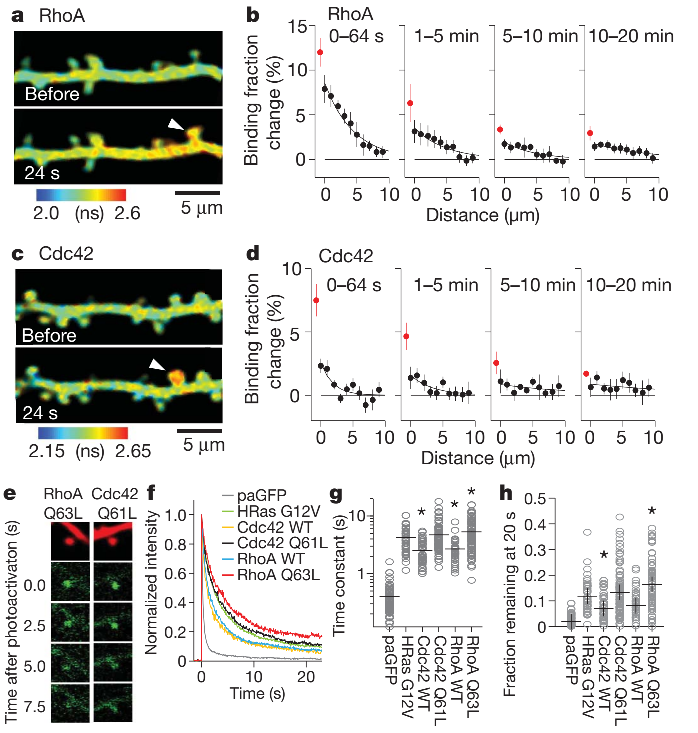


Figure 3 | Spatial profile of RhoA and Cdc42 activities. a, Fluorescence lifetime images of RhoA activity before and after glutamate uncaging. Arrowheads in **a** and **c** indicate the stimulated spine. **b,** Averaged spatial profile of RhoA activation. Red circles indicate the activity in the stimulated spine, and black circles indicate the activity in the dendrite, plotted as a function of the distance along the dendrite from the simulated spine. The number of samples (dendrites/neurons) is 20/18. **c,** Fluorescence lifetime images of Cdc42 activity. **d,** Averaged spatial profile of Cdc42 activation. The number of samples (dendrites/neurons) is 30/26. **e,** The fluorescence images of paGFP–RhoA (left) and paGFP–Cdc42 (right) after spine-head photoactivation (green, paGFP-Rho GTPases; red, tandem mCherry). **f,** Averaged timecourse of fluorescence decay in spines after photoactivation of paGFP-tagged proteins in the spines. The fluorescence intensity was normalized to the peak. The numbers of samples (spines/neurons) are 63/6 for paGFP, 38/4 for paGFP–HRas (G12V), 41/5 for paGFP–Cdc42 (WT), 83/9 for paGFP–Cdc42 (Q61L), 40/4 for paGFP– RhoA (WT) and 79/10 for paGFP–RhoA (Q63L). HRas (G12V), RhoA (Q63L) and Cdc42 (Q61L) are constitutively active mutants. **g, h,** Decay time constants (**g**) and the fraction remaining at 20 s (**h**) of paGFP fluorescence in the photoactivated spines. Horizontal bars indicate the means.

©2011 Macmillan Publishers Limited. All rights reserved

dendrite and decayed sharply with a length constant of around 1.9 μm (Fig. 3c and d). These experiments were performed at room temperature (25–27 °C), but similar results were also obtained at near-physiological temperature (32–34 °C; Supplementary Figs 6 and 7).

To test whether the difference in the degree of the compartmentalization of RhoA and Cdc42 is due to a difference in their mobility[12,13], we measured spine–dendrite diffusion coupling using photoactivatable GFP (paGFP)[20] fused to Rho GTPases. Following photoactivation of paGFP in a spine, the fluorescence intensity in the spine decayed owing to the diffusion of paGFP–Rho GTPases out of the spine with a time constant of about 3 s for the wild type and about 5 s for their constitutively active mutants (Fig. 3e–h). These values are about ten times larger than the decay time constant of cytosolic paGFP (~0.4 s) and similar to that of a constitutively active HRas mutant (~5 s) (Fig. 3e–h)[12]. The difference between wild-type Rho GTPases and constitutively active mutants presumably reflects the difference in the fraction of the protein bound to the plasma membrane, given that active Rho GTPases are localized on the plasma membrane[21]. There was only a small fraction (10–20%) of fluorescence remaining at 20 s after photoactivation (Fig. 3e–h), suggesting that no major immobile fraction of RhoA or Cdc42 exists in spines. Thus, Cdc42 is as mobile as RhoA and HRas, yet only Cdc42 shows the compartmentalized activity.

Next, to elucidate the roles of Rho GTPase activation in spine structural plasticity, we measured the spine volume change under the inhibition of Rho or Cdc42 signalling (Fig. 4a–j). Downregulation of RhoA and RhoB with short-hairpin RNA (shRNA) decreased the transient volume change, but did not appreciably affect the sustained volume change (Fig. 4a, i and j). In contrast, shRNA against Cdc42 decreased the sustained volume change, but not the transient volume change (Fig. 4e, i and j). The phenotypes caused by these shRNAs were rescued by co-expressing shRNA-resistant mEGFP–Rho GTPases (Fig. 4b, f, i and j), indicating that the effect of the shRNAs is specific

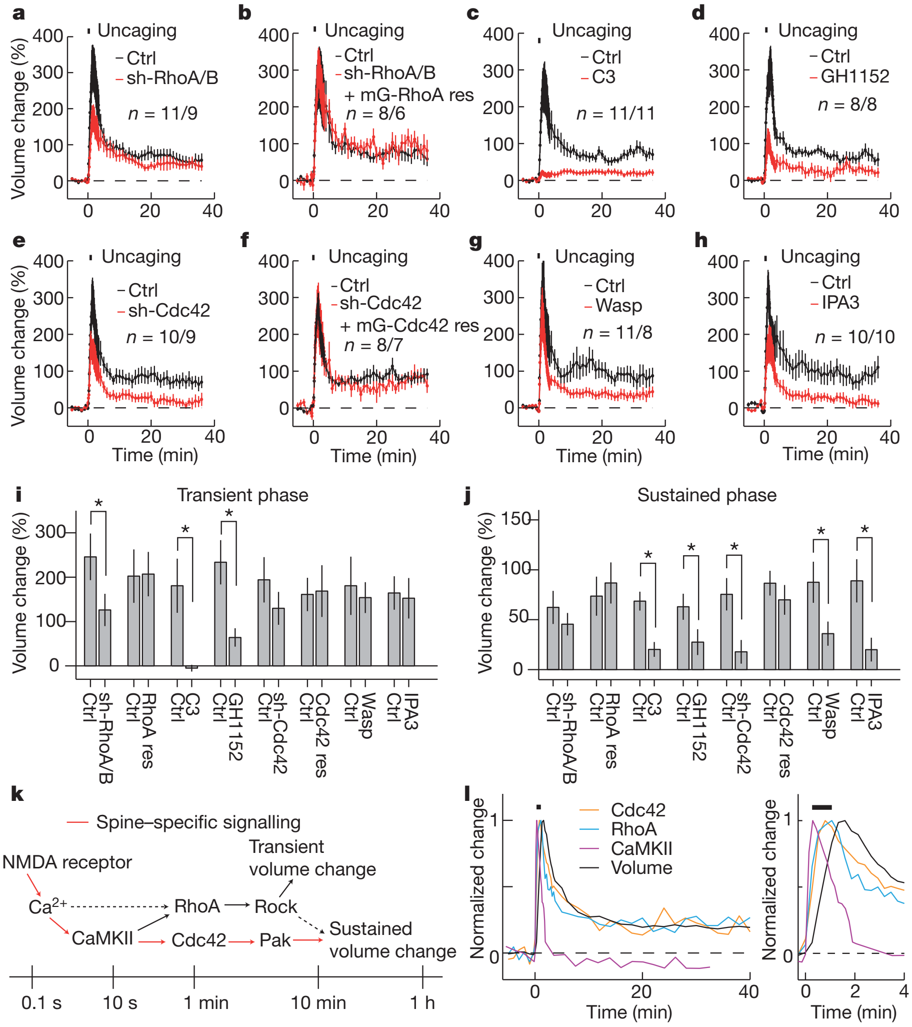


Figure 4 | The effect of Rho GTPase inhibition for structural plasticity of spine head enlargement. a–h, Averaged time course of spine volume change in stimulated spines in neurons under manipulations of Rho GTPase signalling. Red traces: neurons were transfected with shRNAs against RhoA and RhoB (sh-RhoA/B) and mEGFP (**a**), sh-RhoA/B, mEGFP–shRNA resistant RhoA (mEGFP–RhoA res) and tandem mCherry (**b**), mCherry–C3 transferase (C3) and mEGFP (**c**), mEGFP (**d, h**), shRNA against Cdc42 (sh-Cdc42) and mEGFP (**e**), sh-Cdc42, mEGFP–shRNA resistant Cdc42 (mEGFP–Cdc42 res) and tandem mCherry (**f**) or mCherry–Wasp(210–321) (Wasp) and mEGFP (**g**). Black traces: paired control experiments were performed in the same batch of slices using a scrambled shRNA instead of targeted shRNAs (**a, e**), mEGFP alone (**b, f**) or mCherry instead of C3 transferase and Wasp (**c, g**). Pharmacology experiments (**d, h**) were performed before (paired control, black) and 30–40 min (red) after applying drugs to the bath. Fluorescence intensity of mEGFP (**a, c, d, e, g, h**) or tandem mCherry (**b, f**) was used to measure the spine volume change. The numbers of samples (spines/neurons) are indicated in the figures (same numbers for control and experiment groups). **i**, Transient volume change (volume change averaged over 1.5–2 min subtracted by that averaged over 20–36 min). Stars denote statistical significance ($P < 0.05$, paired *t*-test). **j**, Sustained volume change (volume change averaged over 20–36 min). **k**, A model of Cdc42 and RhoA activation. **l**, Superimposed time courses of spine volume change and activation of RhoA (Fig. 1b), Cdc42 (Fig. 2b) and CaMKII[13] in spines undergoing structural plasticity. The time courses were normalized to the peak. The right-hand panel shows a closer view of the first 4 min.

©2011 Macmillan Publishers Limited. All rights reserved

and the mEGFP–RhoA and mEGFP–Cdc42 used in the FRET sensors are functional as endogenous proteins. Because downregulation of proteins with shRNA is partial (Supplementary Fig. 8) and requires a relatively long time (4 days), we also inhibited Rho and Cdc42 signalling by expressing mCherry–C3 transferase, a Rho inhibitor[22], and the Cdc42 binding domain of Wasp (221–321) tagged with mCherry (Wasp)[23], respectively, for shorter time (24 h). Rho inhibition with C3 transferase inhibited both the transient and the sustained phases (Fig. 4c, i and j), showing stronger effects than shRNA (Fig. 4a, i and j). Cdc42 inhibition with Wasp inhibited the sustained phase but not the transient phase (Fig. 4g, i and j), consistent with the shRNA result (Fig. 4e, i and j). Thus, our data suggest that Rho signalling is required for the transient phase and probably the sustained phase of spine enlargement, whereas Cdc42 signalling is required for the sustained phase. Neither C3 transferase nor Wasp affected CaMKII activation (Supplementary Fig. 9), indicating that there is no feedback signalling from Rho and Cdc42 to upstream Ca^{2+} and CaMKII. C3 transferase and Wasp also inhibited synaptic potentiation induced by pairing postsynaptic depolarization (0 mV) and two-photon glutamate uncaging (Supplementary Fig. 10)[13,14], suggesting that Rho and Cdc42 are important for the functional plasticity as well as the structural plasticity of spines.

Among known effectors of Rho and Cdc42, Rock and Pak are two kinases that can be activated respectively by these GTPases[24–26]. We tested whether they are required for structural plasticity through acute (30–40 min) application of specific pharmacological inhibitors. Inhibition of Rock with Glycyl-H1152 (2 μM)[27] suppressed both transient and sustained volume change (Fig. 4d, i and j), similarly to the Rho inhibitor C3 transferase (Fig. 4c). In contrast, inhibition of Pak with inhibitor targeting Pak1 activation-3 (IPA3) (100 μM)[28] decreased the sustained volume change selectively without changing the transient volume change (Fig. 4h, i and j), similarly to inhibition of Cdc42 signalling (Fig. 4e and g). Taken together with the results from Rho/Cdc42 inhibition (Fig. 4a–j), our data implies that the Rho–Rock pathway is required for both the transient and sustained phases, whereas the Cdc42–Pak pathway is required for the sustained phase of the structural plasticity but not for the transient phase (Fig. 4k).

In this study, we visualized RhoA and Cdc42 activation in single dendritic spines undergoing structural plasticity associated with long-term potentiation[12–15,19]. The time course of their activation was similar to that of the volume change: rapid activation was followed by persistent activation lasting more than 30 min (Fig. 4l). As expected from its high mobility (Fig. 3), RhoA spread into the dendrite upon activation (Fig. 1a–c)[13]. However, the activity invasion into adjacent spines was relatively small (25% of the stimulated spines, Fig. 1b and c) and was not sufficient to produce plasticity (Fig. 1d and e). In contrast with the diffusive pattern of RhoA activity, Cdc42 activity was restricted to the stimulated spines (Figs 2 and 3). The compartmentalization of Cdc42 activity is not due to the limited diffusion of active Cdc42, because active Cdc42 is as mobile as RhoA and HRas (Fig. 3e–h). Given that the high spatial gradient of Cdc42 between the stimulated spines and dendrite was maintained for more than 30 min (Fig. 2b and c, and 3d), Cdc42 must be continuously activated at the stimulated spines during plasticity, and inactivated immediately after diffusing out of the spines. The short length constant of Cdc42 in the dendrites (1.9 μm, Fig. 3) also supports the fast inactivation of Cdc42 in the dendrite. The inactivation time constant τ is related to the length constant L and the diffusion constant $D \approx 0.6\ \mu m^2$ (ref. 12) as follows: $\tau = L^2/D$, and so τ is about 6 s for Cdc42 and about 30 s for RhoA, compared to 200–300 s for HRas[12].

Our results further indicated that RhoA and Cdc42 activation is CaMKII-dependent, and activation lasts for more than 30 min (Figs 1 and 2). The previous imaging study suggested that CaMKII activity decays with a time constant of about 10 s (Fig. 4k and l)[13], thereby integrating NMDA-receptor-evoked Ca^{2+} transients that last for about 0.1 s (refs 13 and 29). Localized, persistent activation of RhoA and Cdc42, which peaks between CaMKII activation and the volume change (Fig. 4l), relays the transient CaMKII signalling[13] to long-term spine structural plasticity (Fig. 4k). In particular, because both CaMKII[13,14] and Cdc42 exhibit spine specific activation and are required for the maintenance of plasticity (Fig. 4e, g, i and j), the NMDA receptor–CaMKII–Cdc42–Pak pathway (red line in Fig. 4k) constitutes the spine-specific signal transduction spanning the timescale from less than one second to more than half an hour to cause sustained structural and functional spine plasticity (Fig. 4k and l).

METHODS SUMMARY

Hippocampal cultured slices were prepared from postnatal day 6–7 rats as described[30]. Neurons were sparsely transfected with Rho GTPase FRET sensors using ballistic gene transfer[18] at days *in vitro* 10–14, and imaged 2–4 days after transfection. Rho–GTPase activity was measured using 2pFLIM (green) and spine volume change was monitored by measuring the fluorescence intensity of mCherry–RBD–mCherry (red) in spines using normal two-photon microscopy (Supplementary Fig. 3)[12,13]. Most of the imaging experiments were performed at room temperature (25–27 °C).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 11 March 2010; accepted 13 January 2011.
Published online 20 March 2011.

1. Tashiro, A. & Yuste, R. Regulation of dendritic spine motility and stability by Rac1 and Rho kinase: evidence for two forms of spine motility. *Mol. Cell. Neurosci.* **26,** 429–440 (2004).
2. Luo, L. Rho GTPases in neuronal morphogenesis. *Nature Rev. Neurosci.* **1,** 173–180 (2000).
3. Saneyoshi, T., Fortin, D. A. & Soderling, T. R. Regulation of spine and synapse formation by activity-dependent intracellular signaling pathways. *Curr. Opin. Neurobiol.* **20,** 108–115 (2009).
4. Fortin, D. A. *et al.* Long-term potentiation-dependent spine enlargement requires synaptic Ca^{2+}-permeable AMPA receptors recruited by CaM-kinase I. *J. Neurosci.* **30,** 11565–11575 (2010).
5. O'Kane, E. M., Stone, T. W. & Morris, B. J. Activation of Rho GTPases by synaptic transmission in the hippocampus. *J. Neurochem.* **87,** 1309–1312 (2003).
6. Rex, C. S. *et al.* Different Rho GTPase-dependent signaling pathways initiate sequential steps in the consolidation of long-term potentiation. *J. Cell Biol.* **186,** 85–97 (2009).
7. Asrar, S. *et al.* Regulation of hippocampal long-term potentiation by p21-activated protein kinase 1 (PAK1). *Neuropharmacology* **56,** 73–80 (2009).
8. Wang, H. G. *et al.* Presynaptic and postsynaptic roles of NO, cGK, and RhoA in long-lasting potentiation and aggregation of synaptic proteins. *Neuron* **45,** 389–403 (2005).
9. Nadif Kasri, N. & Van Aelst, L. Rho-linked genes and neurological disorders. *Pflugers Arch.* **455,** 787–797 (2008).
10. Hotulainen, P. & Hoogenraad, C. C. Actin in dendritic spines: connecting dynamics to function. *J. Cell Biol.* **189,** 619–629 (2010).
11. Yasuda, R. *et al.* Supersensitive Ras activation in dendrites and spines revealed by two-photon fluorescence lifetime imaging. *Nature Neurosci.* **9,** 283–291 (2006).
12. Harvey, C. D., Yasuda, R., Zhong, H. & Svoboda, K. The spread of Ras activity triggered by activation of a single dendritic spine. *Science* **321,** 136–140 (2008).
13. Lee, S. J., Escobedo-Lozoya, Y., Szatmari, E. M. & Yasuda, R. Activation of CaMKII in single dendritic spines during long-term potentiation. *Nature* **458,** 299–304 (2009).
14. Matsuzaki, M., Honkura, N., Ellis-Davies, G. C. & Kasai, H. Structural basis of long-term potentiation in single dendritic spines. *Nature* **429,** 761–766 (2004).
15. Honkura, N., Matsuzaki, M., Noguchi, J., Ellis-Davies, G. C. & Kasai, H. The subspine organization of actin fibers regulates the structure and plasticity of dendritic spines. *Neuron* **57,** 719–729 (2008).
16. Heasman, S. J. & Ridley, A. J. Mammalian Rho GTPases: new insights into their functions from *in vivo* studies. *Nature Rev. Mol. Cell Biol.* **9,** 690–701 (2008).
17. Okamoto, K., Nagai, T., Miyawaki, A. & Hayashi, Y. Rapid and persistent modulation of actin dynamics regulates postsynaptic reorganization underlying bidirectional plasticity. *Nature Neurosci.* **7,** 1104–1112 (2004).
18. McAllister, A. K. Biolistic transfection of neurons. *Sci. STKE* **2000,** pl1 (2000).
19. Harvey, C. D. & Svoboda, K. Locally dynamic synaptic learning rules in pyramidal neuron dendrites. *Nature* **450,** 1195–1200 (2007).
20. Patterson, G. H. & Lippincott-Schwartz, J. A photoactivatable GFP for selective photolabeling of proteins and cells. *Science* **297,** 1873–1877 (2002).
21. Michaelson, D. *et al.* Differential localization of Rho GTPases in live cells: regulation by hypervariable regions and RhoGDI binding. *J. Cell Biol.* **152,** 111–126 (2001).
22. Wilde, C., Genth, H., Aktories, K. & Just, I. Recognition of RhoA by *Clostridium botulinum* C3 exoenzyme. *J. Biol. Chem.* **275,** 16478–16483 (2000).
23. Elliot-Smith, A. E., Mott, H. R., Lowe, P. N., Laue, E. D. & Owen, D. Specificity determinants on Cdc42 for binding its effector protein ACK. *Biochemistry* **44,** 12373–12383 (2005).

©2011 Macmillan Publishers Limited. All rights reserved

24. Nikolić, M. The Pak1 kinase: an important regulator of neuronal morphology and function in the developing forebrain. *Mol. Neurobiol.* **37,** 187–202 (2008).
25. Kreis, P. *et al.* The p21-activated kinase 3 implicated in mental retardation regulates spine morphogenesis through a Cdc42-dependent pathway. *J. Biol. Chem.* **282,** 21497–21506 (2007).
26. Iden, S. & Collard, J. G. Crosstalk between small GTPases and polarity proteins in cell polarization. *Nature Rev. Mol. Cell Biol.* **9,** 846–859 (2008).
27. Tamura, M. *et al.* Development of specific Rho-kinase inhibitors and their clinical application. *Biochim. Biophys. Acta* **1754,** 245–252 (2005).
28. Deacon, S. W. *et al.* An isoform-selective, small-molecule inhibitor targets the autoregulatory mechanism of p21-activated kinase. *Chem. Biol.* **15,** 322–331 (2008).
29. Sabatini, B. L., Oertner, T. G. & Svoboda, K. The life cycle of Ca(2+) ions in dendritic spines. *Neuron* **33,** 439–452 (2002).
30. Stoppini, L., Buchs, P. A. & Muller, D. A simple method for organotypic cultures of nervous tissue. *J. Neurosci. Methods* **37,** 173–182 (1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Hahn, G. Bokock, A. Aplin, S. Soderling, M. Matsuda and Y. Hayashi for plasmids and reagents, S.-J. Lee for CaMKII activation data, S. Soderling, S. Raghavachari, L. van Aelst, Y. Hayashi, H. Kasai and M. Ehlers for discussion, and H. Hedrick and M. Patterson for comments on the manuscript. We also thank A. Wan for preparing cultured slices and D. Kloetzer for laboratory management. This study was funded by the Howard Hughes Medical Institute, the National Institute of Mental Health, the National Institute of Neurological Disorders and Stroke, the National Institute of Drug Abuse, the Alzheimer's Association and the Japan Society for the Promotion of Science (H.M.).

Author Contributions H.M. and R.Y. designed the experiments. H.M. performed the experiments and data analysis. H.W. performed electrophysiological experiments. H.M. and R.Y. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.Y. (yasuda@neuro.duke.edu).

©2011 Macmillan Publishers Limited. All rights reserved

METHODS

Plasmids. Plasmids containing Dbl/p50RhoGAP, Rhotekin(8–89), C3 transferase, RhoA/Cdc42/Pak1(65–118), and Wasp were the kind gifts of K. Hahn, G. Bokock, A. Aplin, M. Matsuda and S. Soderling, respectively. Pak3(60–113) was prepared by introducing mutations M99I and S115L into Pak1(65–118). mEGFP–RhoA and mEGFP–Cdc42 were prepared by inserting human RhoA and Cdc42 coding sequences into the pEGFP–C1 vector (Clontech) containing the A206K monomeric mutation in EGFP[31]. Because the transcription of human RhoA is terminated at the codon 349–353 (AATAA) in *Escherichia coli*, we introduced a silent mutation from AATAA to AACAA at the site, which does not change the amino acid sequence. This sequence was used for all experiments. The linker between mEGFP and Rho GTPases (RhoA and Cdc42) is SGLRSRG. Tandem mCherry[32] was prepared by replacing the EGFP of the pEGFP–N3 vector with two mCherry. mCherry–Rhotekin(8–89)–mCherry and mCherry–Pak3(60–113)–mCherry were prepared by inserting Rhotekin(8–89) and Pak3(60–113) into tandem mCherry, respectively. The linkers between Rhotekin(8–89) and mCherry are SGLRSG for the amino terminus and VDVTAGPGSG for the carboxy terminus. The linkers between Pak3(60–113) and mCherry are SGLRSRG for the N terminus and GSG for the C terminus. Photoactivatable GFP (paGFP)[20]–Rho GTPases were prepared by replacing the mEGFP of the mEGFP–Rho GTPases with paGFP (A206K). Mutations were introduced using a Site-Directed Mutagenesis kit (Stratagene). shRNAs were prepared using the pSuper vector (Oligoengine) with kanamycin resistance gene. The following target sequences were used for shRNA (5′–3′): GGGCAAGAGGATTATGACA for Cdc42 (rat and human), GAAGG ATCTTCGGAATGAT for RhoA (mouse, rat and human), CATCTTGGTGG CCAACAAA for RhoB (rat) and GTGTTGAAGTATCTGTACG for control. shRNA-resistant RhoA and Cdc42 were prepared by introducing three silent mutations in the targeted sequences. mCherry–C3 transferase (34-end) and mCherry–Wasp(210–321) were prepared into the pEGFP–C1 vector without EGFP.

Proteins. Polyhistidine-tagged mEGFP, mCherry, super-folder GFP (sfGFP)–Rho GTPases[33], mCherry–RBDs and their mutants were cloned into the pRSET bacterial expression vector (Invitrogen). Proteins were overexpressed in *E. coli* (DH5α), purified with a Ni$^+$-nitrilotriacetate column (HiTrap, GE Healthcare), and desalted with a desalting column (PD10, GE Healthcare) equilibrated with phosphate buffered saline. The concentration of the purified protein was measured by the absorbance of the fluorophore (mEGFP, $A_{489\,\mathrm{nm}} = 56{,}000\,\mathrm{cm}^{-1}\,\mathrm{M}^{-1}$ (ref. 13); sfGFP, $A_{489\,\mathrm{nm}} = 83{,}000\,\mathrm{cm}^{-1}\,\mathrm{M}^{-1}$ (ref. 33); mCherry, $A_{587\,\mathrm{nm}} = 72{,}000\,\mathrm{cm}^{-1}\,\mathrm{M}^{-1}$ (ref. 32)) or Bradford assay.

Preparation. Hippocampal slices were prepared from postnatal day 6–7 rats, as described[30], in accordance with the animal care and use guidelines of Duke University Medical Center. After 1 to 2 weeks in culture, CA1 pyramidal neurons were transfected with ballistic gene transfer[34] using gold beads (8–12 mg) coated with plasmids containing 30 μg of total complementary DNA (donor:acceptor = 1:1), and imaged 2–4 days after transfection.

HeLa and Rat1 cells were cultured in Dulbecco's modified Eagle medium supplemented with 10% fetal calf serum at 37 °C in 5% CO_2 and transfected using Lipofectamine (Invitrogen).

Two-photon fluorescence lifetime imaging. Details of FRET imaging using a custom-built two-photon fluorescence lifetime imaging microscope have been described previously[35,36]. mEGFP and mCherry were simultaneously excited with a Ti:sapphire laser (Maitai, Spectraphysics) tuned at a wavelength of 920 nm. The fluorescence was collected by an objective (60×, numerical aperture 0.9, Olympus), divided with a dichroic mirror (565 nm) and detected with two separated photoelectron multiplier tubes placed after wavelength filters (Chroma, HQ510/70-2p for green and HQ620/90-2p for red). For fluorescence lifetime imaging in the green channel, a photoelectron multiplier tube with low transfer time spread (H7422-40p; Hamamatsu) was used. A wide-aperture photoelectron multiplier tube (R3896; Hamamatsu) was used for the red channel. Fluorescence lifetime images were obtained using a time-correlated single photon counting board (SPC-140; Becker and Hickl) controlled with custom software[11]. The red signal was acquired using a separate data acquisition board (PCI-6110) and Scanimage software[37].

Two-photon glutamate uncaging. A second Ti:sapphire laser tuned at a wavelength of 720 nm was used to uncage 4-methoxy-7-nitroindolinyl-caged-L-glutamate (MNI-caged glutamate) in extracellular solution with a train of 4–6-ms, 8-mW pulses (30 times at 0.5 Hz) near a spine of interest. Experiments were performed in Mg^{2+} fee artificial cerebral spinal fluid (127 mM NaCl, 2.5 mM KCl, 4 mM $CaCl_2$, 25 mM $NaHCO_3$, 1.25 mM NaH_2PO_4 and 25 mM glucose) containing 1 μM tetrodotoxin and 2 mM MNI-caged L-glutamate aerated with 95% O_2 and 5% CO_2 at 25–27 °C, as described previously[13]. In Supplementary Figs 6 and 7, neurons were maintained at 32–34 °C using a temperature controller (TC324B, SW-10/6 and SH-27B, Warner Instruments).

Two-photon photoactivation of photoactivatable GFP. Two-photon images of paGFP tagged proteins were acquired every 64 ms using a Ti:sapphire laser tuned at a wavelength of 940 nm. For photoactivation of paGFP, the uncaging laser tuned at a wavelength of 800 m was used to apply a pulse of 8 mW with 6–10 ms duration at a spine head. To determine the decay time constant τ and the immobile fraction f_{im}, the paGFP fluorescence F was fitted with an exponential function, $F(t) = F_0\exp(-t/\tau) + f_{\mathrm{im}}$, where F_0 is the fluorescence intensity at $t = 0$.

2pFLIM data analyses. To obtain the mEGFP fluorescence lifetime, we summed over all pixels in an image of a cell expressing mEGFP–Rho GTPases, and fitted a fluorescence lifetime curve with a single exponential function convolved with the Gaussian pulse response function:

$$F(t) = F_0 H(t, t_0, \tau_D, \tau_G) \tag{1}$$

where F_0 is the constant, and

$$H(t,t_0,\tau_D,\tau_G) = \frac{1}{2}\exp\left(\frac{\tau_G^2}{2\tau_D} - \frac{t-t_0}{\tau_D}\right)\mathrm{erf}\left(\frac{\tau_G^2 - \tau_D(t-t_0)}{\sqrt{2}\tau_D\tau_G}\right) \tag{2}$$

in which τ_D is the fluorescence lifetime of the free donor (mEGFP–Rho GTPase), τ_G is the width of the Gaussian pulse response function, F_0 is the peak fluorescence before convolution and t_0 is the time offset, and erf is the error function. To measure the fraction of donor bound to acceptor, we summed all pixels over a whole image and fitted a fluorescence lifetime curve with a double exponential function convolved with the Gaussian pulse response function:

$$F(t) = F_0[P_D H(t, t_0, \tau_D, \tau_G) + P_D H(t, t_0, \tau_{AD}, \tau_G)] \tag{3}$$

where τ_{AD} is the fluorescence lifetime of donor bound with acceptor, and P_D and P_{AD} are the fractions of free donor and donor bound with acceptor, respectively. We fixed τ_D to the fluorescence lifetime obtained from free mEGFP–Rho GTpase (2.59 ns). To generate the fluorescence lifetime image, we calculated the mean photon arrival time, $<t>$, in each pixel as[38]:

$$<t> = \int tF(t)\mathrm{d}t/\int F(t)\mathrm{d}t$$

The mean photon arrival time is then related to the mean fluorescence lifetime $<\tau>$ by an offset arrival time t_o, which is obtained by fitting the whole image[35]:

$$<\tau> = <t> - t_0$$

For small regions of interest in an image (spines or dendrites), we calculated the binding fraction (P_{AD}) as[35]:

$$P_{AD} = \tau_D(\tau_D - <\tau>)(\tau_D - \tau_{AD})^{-1}(\tau_D + \tau_{AD} - <\tau>)^{-1} \tag{4}$$

Overexpression level of Rho GTPase sensors in neurons. The concentration of mEGFP–Rho GTPase and mCherry–RBD–mCherry in neurons was estimated by measuring fluorescence intensity of mEGFP and mCherry in thick apical dendrites under two-photon microscopy relative to that of purified, polyhistidine-tagged mEGFP (1 μM) and mCherry (10 μM), respectively[39].

Measurements of the affinity between Rho GTPases and RBDs. Purified sfGFP–Rho GTPases (RhoA and Cdc42) were loaded with GppNHp (2′,3′-O-*N*-methyl anthraniloyl–GppNHp) and GDP by incubating in the presence of tenfold molar excess of GppNHp and GDP in $MgCl_2$-free phosphate buffered saline containing 1 mM EDTA for 10 min, respectively. The reaction was terminated by adding 10 mM $MgCl_2$ (ref. 40). sfGFP–Rho GTPases and mCherry–RBD were mixed and incubated at room temperature for 20 min. FRET between sfGFP and mCherry was measured under 2pFLIM, and the fraction of sfGFP–Rho GTPases bound to mCherry–RBD was calculated by fitting the fluorescence lifetime curve with a double exponential function (equation (3)). The dissociation constant was obtained by fitting the relationship between the binding fraction and the concentration of mCherry–RBD with a Michaelis–Menten function (Supplementary Fig. 1).

Estimation of endogenous Rho GTPase concentration. We determined the concentrations of CaMKIIα, RhoA, and Cdc42 in the CA1 region of hippocampal slice culture by semiquantitative western blotting (Supplementary Fig. 15). First, the CA1 regions from ten slices were collected and weighed. The series of purified CaMKIIα (4 μM, 10 μM, 20 μM, 30 μM, 40 μM), RhoA (0.1 μM, 0.2 μM, 0.5 μM, 1.0 μM, 1.5 μM), or Cdc42 (0.1 μM, 0.2 μM, 0.5 μM, 1.0 μM, 1.5 μM) were prepared to the same weights as the CA1 tissue. The CA1 tissue and purified proteins were dissolved in SDS sample buffer, and analysed by western blotting. The following antibodies were used: anti-CaMKII (EP1829Y; Abcam); anti-RhoA (26C4; Santa Cruz Biotechnology); anti-Cdc42 (BD44; BD Transduction Laboratories); goat anti-mouse (Zymax). Chemiluminescence signals were detected using a Storm image acquisition system (Molecular Dynamics), and analysed digitally using ImageJ software. By comparing the band intensities of the purified proteins

©2011 Macmillan Publishers Limited. All rights reserved

to that of lysates from the CA1 tissue, we calculated the concentration of the respective proteins in the CA1 tissue (Supplementary Fig. 6). The concentration was estimated to be 23.4 μM for CaMKIIα, 0.38 μM for RhoA and 0.25 μM for Cdc42. The estimation of CaMKIIα concentration is consistent with that obtained from immunofluorescence[13].

Electrophysiology. Whole-cell patch clamping was performed with patch pipettes (4–9 MΩ) containing Cs^+ internal solution (130 mM $CsMeSO_3$, 10 mM Na-phosphocreatine, 4 mM $MgCl_2$, 4 mM Na_2-ATP, 0.4 mM Na_2-GTP, 10 mM Cs-HEPES [pH 7.3])[13]. Excitatory postsynaptic current evoked by two-photon glutamate uncaging at a spine was measured through the patch pipette using a patch-camp amplifier (Multiclamp 700B, Molecular Devices). Long-term potentiation was induced within 5 min of patching by pairing depolarization (0 mV) with two-photon glutamate uncaging at a spine (0.5 Hz, 4 ms, 30 pulses). Neurons showing more than 20% drift in the input or series resistances were not used for further analyses.

31. Zacharias, D. A., Violin, J. D., Newton, A. C. & Tsien, R. Y. Partitioning of lipid-modified monomeric GFPs into membrane microdomains of live cells. *Science* **296,** 913–916 (2002).
32. Shaner, N. C. *et al.* Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nature Biotechnol.* **22,** 1567–1572 (2004).
33. Pédelacq, J. D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nature Biotechnol.* **24,** 79–88 (2005).
34. O'Brien, J. A. & Lummis, S. C. Biolistic transfection of neuronal cultures using a hand-held gene gun. *Nature Protocols* **1,** 977–981 (2006).
35. Yasuda, R. Imaging spatiotemporal dynamics of neuronal signaling using fluorescence resonance energy transfer and fluorescence lifetime imaging microscopy. *Curr. Opin. Neurobiol.* **16,** 551–561 (2006).
36. Murakoshi, H., Lee, S. J. & Yasuda, R. Highly sensitive and quantitative FRET-FLIM imaging in single dendritic spines using improved non-radiative YFP. *Brain Cell Biol.* **36,** 31–42 (2008).
37. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2,** 13 (2003).
38. Lakowicz, J. R. *Principles of Fluorescence Spectroscopy* (Plenum, 2006).
39. Lee, S. J. & Yasuda, R. Spatiotemporal regulation of signaling in and out of dendritic spines. *CaMKII and Ras Open Neurosci. J.* **3,** 117–127 (2010).
40. Zhao, J., Wang, W. N., Tan, Y. C., Zheng, Y. & Wang, Z. X. Effect of Mg(2+) on the kinetics of guanine nucleotide binding and hydrolysis by Cdc42. *Biochem. Biophys. Res. Commun.* **297,** 653–658 (2002).

©2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09848

Dampening of death pathways by schnurri-2 is essential for T-cell development

Tracy L. Staton[1], Vanja Lazarevic[1], Dallas C. Jones[1], Amanda J. Lanser[1], Tsuyoshi Takagi[2], Shunsuke Ishii[2] & Laurie H. Glimcher[1,3,4]

Generation of a diverse and self-tolerant T-cell repertoire requires appropriate interpretation of T-cell antigen receptor (TCR) signals by CD4^{+} CD8^{+} double-positive thymocytes. Thymocyte cell fate is dictated by the nature of TCR–major-histocompatibility-complex (MHC)–peptide interactions, with signals of higher strength leading to death (negative selection) and signals of intermediate strength leading to differentiation (positive selection)[1]. Molecules that regulate T-cell development by modulating TCR signal strength have been described but components that specifically define the boundaries between positive and negative selection remain unknown. Here we show in mice that repression of TCR-induced death pathways is critical for proper interpretation of positive selecting signals *in vivo*, and identify schnurri-2 (Shn2; also known as Hivep2) as a crucial death dampener. Our results indicate that *Shn2*$^{-/-}$ double-positive thymocytes inappropriately undergo negative selection in response to positive selecting signals, thus leading to disrupted T-cell development. *Shn2*$^{-/-}$ double-positive thymocytes are more sensitive to TCR-induced death *in vitro* and die in response to positive selection interactions *in vivo*. However, Shn2-deficient thymocytes can be positively selected when TCR-induced death is genetically ablated. *Shn2* levels increase after TCR stimulation, indicating that integration of multiple TCR–MHC–peptide interactions may fine-tune the death threshold. Mechanistically, Shn2 functions downstream of TCR proximal signalling compenents to dampen Bax activation and the mitochondrial death pathway. Our findings uncover a critical regulator of T-cell development that controls the balance between death and differentiation.

Cell intrinsic defects in the generation of T cells generally map to either TCR proximal signalling molecules or to two major pathways: the calcineurin/NFAT and Raf/Mek/Erk pathways[2–5]. Shn family proteins are large zinc-finger-containing proteins that regulate morphogenesis in *Caenorhabditis elegans* and *Drosophila*[6,7]. In vertebrates, they influence bone formation, adipogenesis and memory T-cell survival[8–10]. Interestingly, *Shn2*$^{-/-}$ mice have very few post-selection double-positive thymocytes or mature single-positive thymocytes but how Shn2 regulates T-cell development has remained a mystery[11] (Fig. 1a).

We began investigating this severe defect by examining TCR proximal signalling events that are known to be required for, or associated with, positive selection. *Shn2*$^{+/+}$ and *Shn2*$^{-/-}$ double-positive thymocytes similarly upregulated CD69, a canonical activation marker, and had comparable patterns of tyrosine phosphorylated proteins after TCR stimulation, indicating that TCR proximal signalling is unaltered (Supplementary Fig. 1). *Shn2*$^{-/-}$ double-positive thymocytes also had equivalent Erk1/2 phosphorylation after TCR activation (Fig. 1b). In addition, *Shn2*$^{-/-}$ double-positive thymocytes showed normal NFATc1 dephosphorylation, NFATc1 nuclear localization, and induction of the NFAT target gene *Egr2* in response to TCR stimulation, thus indicating that the calcineurin/NFAT pathway is activated normally in these cells (Fig. 1c and Supplementary Fig. 2, data not shown).

In further support of this, microarray analysis of CD4^{+} CD8^{+} CD69^{-} thymocytes before and after TCR stimulation yielded very few differences in basal and induced gene expression between *Shn2*$^{+/+}$ and *Shn2*$^{-/-}$ thymocytes (data not shown). A second, independent microarray analysis also did not detect significant changes in gene expression (T.T. and S.I., unpublished observations). To further examine specific transcription factor activity we transfected luciferase reporter constructs into control and *Shn2*$^{-/-}$ primary double-positive thymocytes. NFAT, AP1, NF-κB and Smad activation was comparable between *Shn2*$^{+/+}$ and *Shn2*$^{-/-}$ double-positive thymocytes (Supplementary Fig. 2). Although evidence in other cell types indicates that Shn proteins may influence NF-κB or Smad activity these pathways are not regulated by Shn2 in thymocytes[8,12].

On the basis of the unexpected finding that Shn2 does not regulate proximal TCR signalling or positive selection pathways, we examined *Shn2*$^{-/-}$ double-positive cells more generally. Thymocyte death *in vivo* can result from either negative selection (in response to strong TCR signals) or death by neglect (in response to weak or no signal)[1]. Therefore, we investigated the response of *Shn2*$^{-/-}$ double-positive thymocytes to a gradient of TCR stimulation. *Shn2*$^{-/-}$ double-positive thymocytes were markedly sensitive to death induced by CD3/CD28 ligation, exemplified by a nearly tenfold shift in responsiveness (Fig. 1d). In contrast, *Shn2*$^{+/+}$ and *Shn2*$^{-/-}$ double-positive thymocytes died comparably in response to CD3 alone, under *in vitro* conditions that approximate death by neglect, and in response to the non-TCR death signals dexamethasone, Fas, or TNF-α treatment (Fig. 1e, f and Supplementary Fig. 3). Thus, although *Shn2*$^{-/-}$ double-positive thymocytes respond normally to non-TCR death signals, they appear to be hyper-responsive to TCR/costimulation-induced cell death. *Shn2*$^{-/-}$ double-positive thymocytes showed increased levels of early death markers, namely caspase activation and disruption of mitochondrial membrane potential, as well as increased upregulation of the late death marker annexin V in response to CD3/CD28 stimulation (Fig. 1d, g). Collectively, these data indicate that *Shn2*$^{-/-}$ double-positive thymocytes do not have an inherent survival defect, but rather have an altered perception of TCR signal strength that leads to inappropriate apoptosis via the intrinsic death pathway.

Because TCR-induced death is specifically affected in *Shn2*$^{-/-}$ double-positive thymocytes, we examined components upstream of the mitochondria in the intrinsic death pathway to map the stage at which Shn2 functions. In thymocytes, TCR signalling can induce Bim and ultimately lead to conversion of Bax to an active conformation, disruption of the mitochondrial membrane, and apoptosis[13–15]. Using a conformation-specific antibody we found increased Bax activation in *Shn2*$^{-/-}$ double-positive thymocytes after TCR stimulation, indicating that increased cell death in *Shn2*$^{-/-}$ double-positive thymocytes is mediated by this pro-apoptotic protein[16] (Fig. 1h and Supplementary Fig. 4).

Our findings indicating that Shn2 deficiency alters the threshold of TCR-mediated cell death led us to consider a role for Shn2 in the balance of positive and negative selection. Strong negative selecting signals activate both the differentiation pathway and the death pathway;

[1]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [2]Laboratory of Molecular Genetics, RIKEN Tsukuba Institute, Tsukuba, Ibaraki 305-0074, Japan. [3]Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]The Ragon Institute of MGH/MIT and Harvard, Charlestown, Massachusetts 02129, USA.

©2011 Macmillan Publishers Limited. All rights reserved

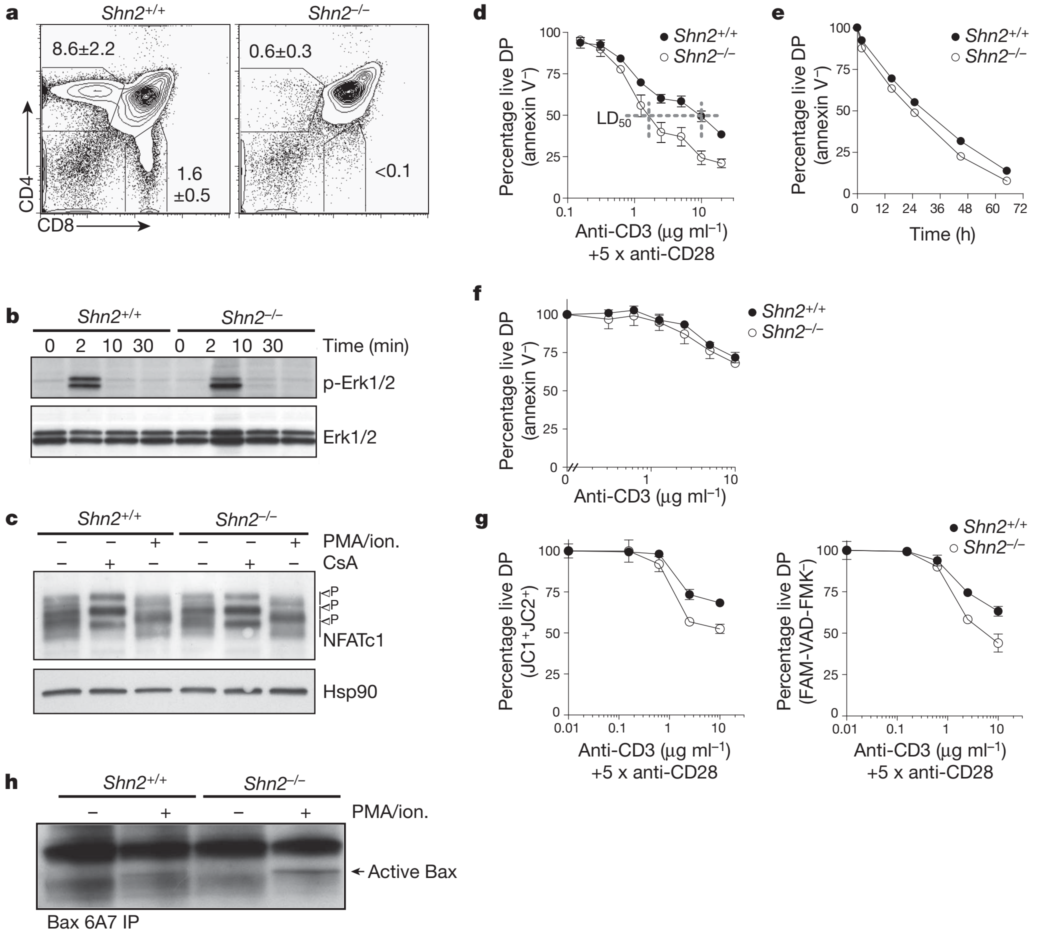


Figure 1 | Increased TCR-induced death in *Shn2*$^{-/-}$ thymocytes. **a**, Representative plots of total thymocytes (mean ± standard deviation, $n > 10$). **b**, Immunoblot of phosphorylated and total Erk1/2 in double-positive thymocytes after CD3 crosslinking. **c**, Immunoblot of NFATc1 in double-positive thymocytes after PMA/ionomycin or CsA treatment for 30 min. **d**–**f**, Double-positive (DP) thymocytes stimulated with plate-bound anti-CD3/anti-CD28 (**d**) or anti-CD3 for 20 h (**f**), or left unstimulated for indicated times (**e**). **g**, Double-positive thymocytes stimulated with plate-bound anti-CD3/anti-CD28 for 20 h. Percentage live double-positive determined by JC1/2 or FAM-VAD-FMK staining. **h**, Immunoblot of total Bax in double-positive thymocytes after 4 h PMA/ionomycin stimulation and immunoprecipitation (IP) with 6A7 antibody.

however, because the death pathway acts in a dominant manner, negative selecting signals result in apoptosis rather than differentiation[2]. In contrast, signals of intermediate strength are able to induce differentiation and positive selection without triggering activation of the intrinsic cell death pathway. It remains unclear how positive selection prevails in response to signals of intermediate strength. We proposed that Shn2 acts to dampen death induced by TCR signalling, thereby allowing positive selection to proceed in response to signals of intermediate strength. In this scenario, *Shn2*$^{-/-}$ thymocytes inappropriately activate death pathways in response to positive selection signals (Supplementary Fig. 12). Therefore, we asked whether preventing *Shn2*$^{-/-}$ thymocytes from dying would allow them to complete positive selection and mature into CD4 and CD8 single-positive cells. The pro-apoptotic protein Bim is a key component of TCR-induced death in thymocytes[17]. We therefore used the *Bim*$^{-/-}$ genetic background as a tool to eliminate TCR-induced death. Genetic inhibition of negative selection by Bim deficiency rescued differentiation of *Shn2*$^{-/-}$ thymocytes, as evidenced by the presence of mature single-positive cells in the thymus and mature T cells in the periphery (Fig. 2a, b and

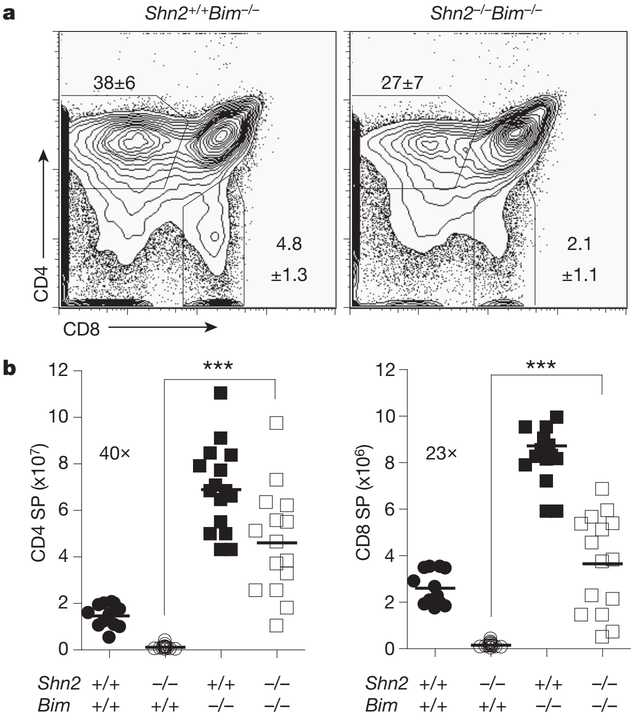
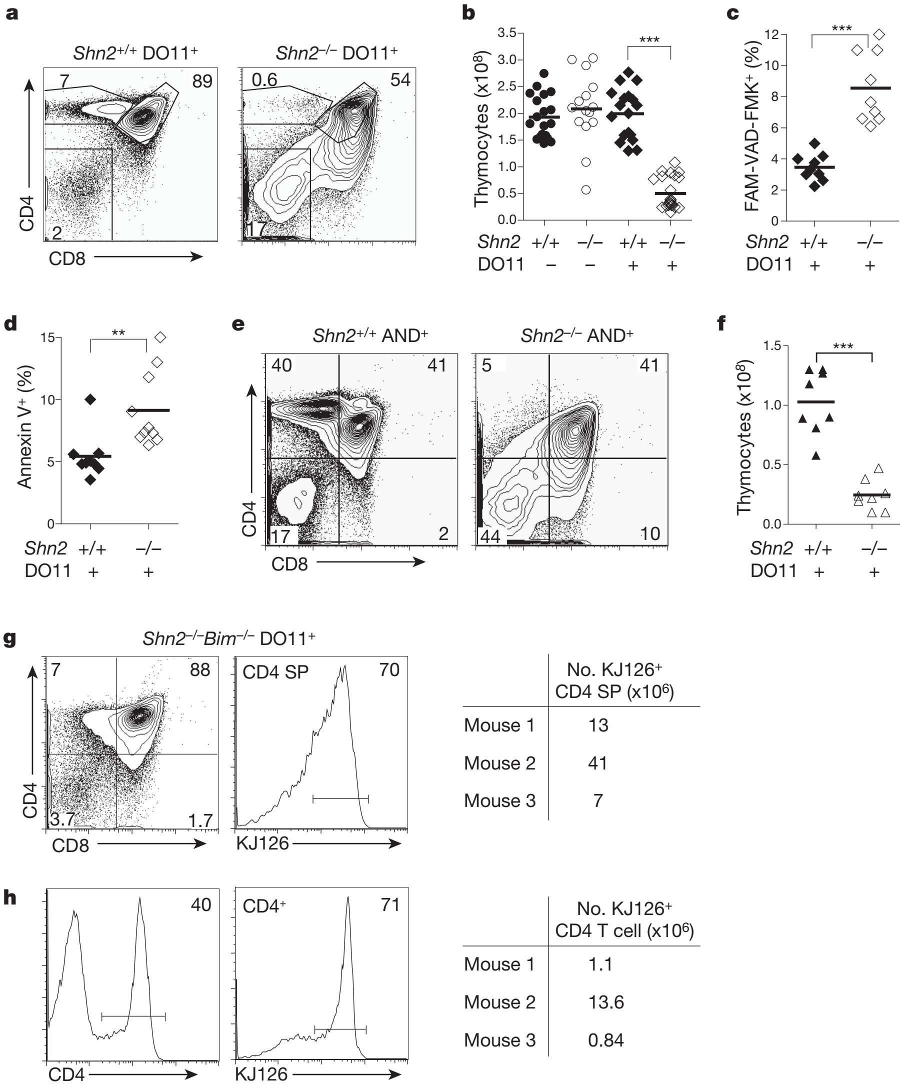


Figure 2 | Bim deficiency rescues positive selection in *Shn2*$^{-/-}$ mice. **a**, Representative plots of total thymocytes (mean ± standard deviation, $n > 10$). **b**, Total numbers of CD4 single-positive (SP) and CD8 single-positive cells from indicated genotypes were calculated to include only mature cells expressing high levels of TCR. Each dot represents a mouse, bar represents the mean, number indicates fold change. $^{***}P < 0.0001$.

©2011 Macmillan Publishers Limited. All rights reserved

Supplementary Fig. 5). These data reinforced the conclusion that differentiation pathways are functional in Shn2-deficient cells, which can mature when death pathways are eliminated.

Next we closely analysed the consequence of positive selection interactions *in vivo* in the presence or absence of Shn2. Our *in vitro* data indicate that in the absence of Shn2 there could be increased death in response to positive selection interactions. However, *Shn2*$^{+/+}$ and *Shn2*$^{-/-}$ mice with polyclonal TCR repertoires have similar total thymocyte numbers and an equally low percentage of dying thymocytes. This may result from the fact that in a diverse, polyclonal repertoire only a small fraction of double-positive thymocytes are undergoing a positive selection at any snapshot in time (Fig. 3b). On the other hand, the majority of double-positive thymocytes in TCR transgenic mice express a TCR of defined specificity; in this system, the effect of positive and negative selecting signals can be dissected and analysed separately. To examine the role of Shn2 in the interpretation of TCR signals in response to positive selection interactions *in vivo*, we crossed *Shn2*$^{-/-}$ mice to three different TCR transgenic strains: DO11, AND and HY. *Shn2*$^{-/-}$ DO11 mice had markedly reduced total and double-positive thymocyte numbers and an altered CD4/CD8 staining profile (Fig. 3a, b). A change in double-positive:double-negative ratio can be indicative of either a developmental block at the double-negative stage or increased double-positive cell death. The decrease in thymocyte number was not due to an early double-negative cell defect, as double-negative cell development was normal in *Shn2*$^{-/-}$ mice (Supplementary Fig. 6). A higher percentage of *Shn2*$^{-/-}$ DO11 thymocytes were annexin V^{+} and showed caspase activation indicating that the decrease in thymic cellularity is due to increased cell death. Hence, we conclude that Shn2 actively represses thymocyte death in response to positive selection interactions *in vivo* (Fig. 3c, d). Results from the AND and the HY TCR transgenic models also indicated that in the absence of Shn2 positive selection was converted to death *in vivo*. *Shn2*$^{-/-}$ AND and *Shn2*$^{-/-}$ HY female mice had reduced thymocyte number and an altered CD4/CD8 staining profile (Fig. 3e, f and Supplementary Fig. 7). We also used the DO11 TCR transgenic system to more closely investigate the rescued single-positive cells in *Shn2*$^{-/-}$ *Bim*$^{-/-}$ mice and determine whether cells could be rescued from death in response to positive selection interactions. The appearance of mature single-positive thymocytes and large numbers of mature T cells in the periphery in *Shn2*$^{-/-}$ *Bim*$^{-/-}$ DO11



	No. KJ126+ CD4 SP (x10^6)
Mouse 1	13
Mouse 2	41
Mouse 3	7

	No. KJ126+ CD4 T cell (x10^6)
Mouse 1	1.1
Mouse 2	13.6
Mouse 3	0.84

Figure 3 | Conversion of positive selection to death *in vivo*. **a**, Representative plots of total thymocytes ($n > 10$). **b**, Total thymocyte cell numbers. **c**, **d**, Percentage of thymocytes (with CD4 single-positive cells excluded) staining with FAM-VAD-FMK reagent (**c**) and annexin V reagent (**d**). **e**, Representative plots of total thymocytes ($n = 8$). **f**, Total thymocyte cell numbers. **g**, **h**, Representative plots of thymus (**g**) and lymph node (**h**) of *Shn2*$^{-/-}$ *Bim*$^{-/-}$ DO11^{+} mice (mouse 3 is shown) with total clonotypic CD4 cell numbers from each mouse shown in the table. For all plots, numbers indicate percentage of cells within each gate. For all graphs, each dot represents a mouse, bar represents the mean. ***$P < 0.0001$, **$P < 0.001$.

©2011 Macmillan Publishers Limited. All rights reserved

TCR transgenic mice indicated that Bim deficiency rescued *Shn2*$^{-/-}$ thymocytes receiving positive selection interactions *in vivo* (Fig. 3g, h and Supplementary Fig. 8).

We interrogated negative selecting interactions *in vivo* in the absence of Shn2. Shn2-deficiency did not alter negative selection in response to the male-specific self antigen in male HY mice or in response to endogenous deleting superantigens in AND mice[18] (Supplementary Fig. 7). Thus, in the absence of Shn2, physiological negative selection is unaltered.

To determine whether TCR–MHC interactions were required to kill *Shn2*$^{-/-}$ DO11 thymocytes *in vivo*, bone marrow from *Shn2*$^{+/+}$ DO11 and *Shn2*$^{-/-}$ DO11 mice was transplanted into irradiated MHC I$^{-/-}$II$^{-/-}$ and wild-type recipients. Thymocyte numbers in MHC I$^{-/-}$II$^{-/-}$ mice reconstituted with *Shn2*$^{-/-}$ DO11 bone marrow were rescued to the level found in MHC I$^{-/-}$II$^{-/-}$ chimeras generated with *Shn2*$^{+/+}$ DO11 bone marrow, whereas thymocyte numbers in wild-type mice that received *Shn2*$^{-/-}$ DO11 bone marrow were still low (Fig. 4a, b). These data, together with the rescue of CD4 single-positive cells in *Shn2*$^{-/-}$ *Bim*$^{-/-}$ DO11 TCR transgenic mice, point to a role for positively selecting interactions killing *Shn2*$^{-/-}$ thymocytes *in vivo*. Additionally, naive *Shn2*$^{-/-}$ DO11 double-positive thymocytes that developed in MHC I$^{-/-}$II$^{-/-}$ chimaeras were more sensitive to death induced by antigen presenting cells (APCs) loaded with OVA peptide *in vitro* as evidenced by enhanced death at lower concentrations of peptide (Fig. 4c). Similar results were obtained when *Shn2*$^{-/-}$ AND double-positive thymocytes were exposed to their cognate antigen *in vitro* (Fig. 4d).

Interestingly, *Shn2* heterozygous mice have only half the normal number of mature single-positive cells, indicating that thymocyte viability is highly sensitive to *Shn* gene dosage *in vivo*[11,19] (Supplementary Fig. 9). Given that thymocytes undergo multiple interactions before being relegated to death or differentiation[20,21], we investigated whether *Shn2* expression levels are static or dynamic during thymocyte development. TCR stimulation or pharmacological activation increases *Shn2* expression by double-positive thymocytes (Fig. 4e–g). Although cyclosporin A (CsA) treatment has a small effect on *Shn2* induction, Ca^{2+}/calmodulin blockade completely inhibits *Shn2* induction, indicating that another calmodulin-dependent pathway regulates TCR-induced *Shn2* expression. The dynamic changes in *Shn2* levels indicate that fine tuning of the death threshold may rely on the integration of previous TCR–MHC interactions and that upregulation of Shn2 expression may serve to adjust death pathway thresholds and allow positive selection of appropriate cells.

During T-cell development the proper interpretation of TCR signals is the critical event that induces the differentiation of productive cells and the elimination of potentially destructive ones. There is a clearly defined border between positive and negative selection that is dependent on the affinity of TCR–MHC interactions[22]. Where and how the signals activating positive and negative selection pathways diverge remains unknown. Here we show that Shn2 is critical in maintaining

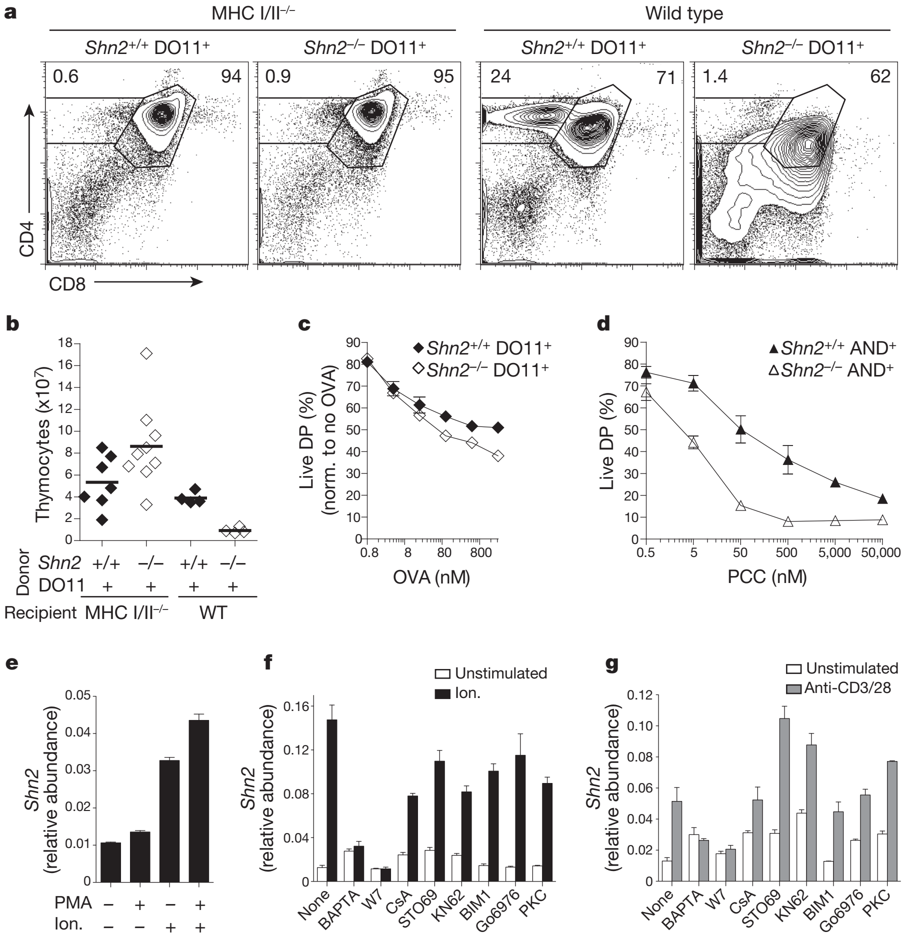


Figure 4 | Shn2 dampens TCR-induced death *in vivo*. a, Representative plots of total thymocytes from bone marrow chimaeras. Numbers indicate percentage of cells within each gate. **b**, Total thymocyte cell numbers. Each dot represents a mouse, bar represents the mean. **c**, Double-positive thymocytes stimulated by M12 cells pulsed with OVA peptide for 20 h. Data is normalized to stimulation with M12 cells alone. **d**, Double-positive thymocytes were stimulated by PI39 cells pulsed with PCC peptide for 20 h. **e–g**, Quantitative PCR analysis of Shn2 expression by double-positive thymocytes after 3 h stimulation with PMA and/or ionomycin (**e**), ionomycin (**f**), or plate-bound anti-CD3/anti-CD28 (**g**). Cells were pretreated with inhibitors for 30 min as indicated.

©2011 Macmillan Publishers Limited. All rights reserved

the balance between death and differentiation by repressing TCR-induced death pathways. Investigation of events upstream of Bax activation revealed that basal and TCR-induced levels of Bim and Nur77, two critical regulators of TCR-induced cell death during negative selection[17,23], were unperturbed in the absence of Shn2 (Supplementary Fig. 10). In addition, phosporylation of Bim, critical for cell death *in vivo*, was unaltered in *Shn2*$^{-/-}$ double-positive thymocytes[24] (Supplementary Fig. 4). Total protein levels of other apoptotic regulatory molecules (Bcl2, Bcl-$_{\mathrm{XL}}$ (also known as Bcl2l1), Mcl1 and Bax) were also normal (Supplementary Fig. 10). These data indicated that Shn2 does not simply regulate the balance between established pro-death and anti-death molecules. Considerable effort has focused on uncovering the intricate mechanism of action by which death proteins regulate cell death in many cell types[25]. Shn2 may control activity of these molecules by regulating their associations with one another or affecting other post-translational modifications. Understanding the exact biochemical mechanism by which Shn2 regulates Bax activation may reveal how the intrinsic death pathway is regulated.

Interestingly, whereas some mature thymocytes do develop in *Shn2*$^{-/-}$ mice there is a complete absence of T cells in *Shn2*$^{-/-}$ *Shn3*$^{-/-}$ mice, indicating a functional redundancy among Shn family members (Supplementary Fig. 11). On the basis of the studies described here, we propose a new model in which the previously enigmatic Shn family of proteins functions to oppose activation of negative selection pathways downstream of TCR–MHC interactions of intermediate strength. In this model, Shn-mediated dampening of the intrinsic death pathway is a requisite component to allow the differentiation of mature T cells.

METHODS SUMMARY

Cell isolation and flow cytometry. Double-positive thymocytes were isolated by positive selection of total thymocytes with anti-CD8α magnetic beads (Miltenyi Biotech) with purity greater than 90%. Antibodies used for flow cytometry were from BD Pharmingen and BioLegend. Death was analysed with FAM Poly Caspases Assay Kit (Molecular Probes) and annexin V and JC-1 Mitoscreen Kits (BD Pharmingen) according to the manufacturer's protocol.

Cell stimulation. For plate-bound CD3/28 stimulation, plates were coated overnight with indicated concentrations of antibodies. *In vitro* death assays were performed in triplicate in 96-well flat-bottom plates with 10^5 cells per well for 20 h. Percentage live double-positive was determined by annexin V staining and normalized to unstimulated values. Antigen specific stimulation: 5×10^5 double-positive thymocytes were co-cultured with 5×10^4 APCs (M12.4.1 cells[26] for DO11 TCR transgenic strains and PI39 cells[27] for AND TCR transgenic strains) and indicated concentrations of $\mathrm{OVA}_{323-339}$ peptide (GenScript) or PCC_{88-104} peptide (Anaspec) for 20 h. Percentage live double-positive was determined by annexin V staining and normalized to values of M12 or P139 cells alone. For RNA and protein analyses, 0.5–1×10^7 cells were stimulated in 6-well plates. For inhibitor studies, cells were stimulated with PMA (5 nM), ionomycin (1 μM), or plate-bound anti-CD3 (10 μg ml^{-1})/anti-CD28 (50 μg ml^{-1}). Cells were pretreated with inhibitors for 30 min: BAPTA-AM (20 μM), CsA (200 ng ml^{-1}), W7 (100 μM), KN62 (10 μM), STO69 (10 μg ml^{-1}), BIM1 (1 μM), Go6976 (1 μM), UO126 (10 μM) and SP600125 (10 μM).

Bax immunoprecipitation. Cells (2×10^7) were stimulated with PMA (5 nM) and ionomycin (1 μM) for 4 h, washed once with PBS, and resuspended in lysis buffer (10 mM HEPES, 150 mM NaCl) containing either 0.2% NP-40 or 1% CHAPs plus protease and phosphatase inhibitors. Active conformation Bax antibody (6A7, eBioscience) was used to immunoprecipitate overnight. Membranes were blotted with rabbit polyclonal Bax antibody (Millipore).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 December 2010; accepted 18 January 2011.

1. Starr, T. K., Jameson, S. C. & Hogquist, K. A. Positive and negative selection of T cells. *Annu. Rev. Immunol.* **21,** 139–176 (2003).
2. Gallo, E. M. *et al.* Calcineurin sets the bandwidth for discrimination of signals during thymocyte development. *Nature* **450,** 731–735 (2007).
3. Neilson, J. R., Winslow, M. M., Hur, E. M. & Crabtree, G. R. Calcineurin B1 is essential for positive but not negative selection during thymocyte development. *Immunity* **20,** 255–266 (2004).
4. Alberola-Ila, J., Forbush, K. A., Seger, R., Krebs, E. G. & Perlmutter, R. M. Selective requirement for MAP kinase activation in thymocyte differentiation. *Nature* **373,** 620–623 (1995).
5. Fischer, A. M., Katayama, C. D., Pagès, G., Pouysségur, J. & Hedrick, S. M. The role of Erk1 and Erk2 in multiple stages of T cell development. *Immunity* **23,** 431–443 (2005).
6. Arora, K. *et al.* The *Drosophila schnurri* gene acts in the Dpp/TGFβ signaling pathway and encodes a transcription factor homologous to the human MBP family. *Cell* **81,** 781–790 (1995).
7. Liang, J. *et al.* The *Caenorhabditis elegans schnurri* homolog *sma-9* mediates stage- and cell type-specific responses to DBL-1 BMP-related signaling. *Development* **130,** 6453–6464 (2003).
8. Jin, W. *et al.* Schnurri-2 controls BMP-dependent adipogenesis via interaction with Smad proteins. *Dev. Cell* **10,** 461–471 (2006).
9. Jones, D. C. *et al.* Regulation of adult bone mass by the zinc finger adapter protein Schnurri-3. *Science* **312,** 1223–1227 (2006).
10. Kimura, M. Y. *et al.* Schnurri-2 controls memory Th1 and Th2 cell numbers *in vivo. J. Immunol.* **178,** 4926–4936 (2007).
11. Takagi, T., Harada, J. & Ishii, S. Murine Schnurri-2 is required for positive selection of thymocytes. *Nature Immunol.* **2,** 1048–1053 (2001).
12. Kimura, M. Y. *et al.* Regulation of T helper type 2 cell differentiation by murine Schnurri-2. *J. Exp. Med.* **201,** 397–408 (2005).
13. Cante-Barrett, K., Gallo, E. M., Winslow, M. M. & Crabtree, G. R. Thymocyte negative selection is mediated by protein kinase C- and Ca^{2+}-dependent transcriptional induction of Bim. *J. Immunol.* **176,** 2299–2306 (2006).
14. Strasser, A. The role of BH3-only proteins in the immune system. *Nature Rev. Immunol.* **5,** 189–200 (2005).
15. Rathmell, J. C., Lindsten, T., Zong, W.-X., Cinalli, R. M. & Thompson, C. B. Deficiency in Bak and Bax perturbs thymic selection and lymphoid homeostasis. *Nature Immunol.* **3,** 932–939 (2002).
16. Hsu, Y. T. & Youle, R. J. Nonionic detergents induce dimerization among members of the Bcl-2 family. *J. Biol. Chem.* **272,** 13829–13834 (1997).
17. Bouillet, P. *et al.* BH3-only Bcl-2 family member Bim is required for apoptosis of autoreactive thymocytes. *Nature* **415,** 922–926 (2002).
18. Kisielow, P., Bluthmann, H., Staerz, U. D., Steinmetz, M. & von Boehmer, H. Tolerance in T-cell-receptor transgenic mice involves deletion of nonmature $\mathrm{CD4^{+}8^{+}}$ thymocytes. *Nature* **333,** 742–746 (1988).
19. Jones, D. C. *et al.* Uncoupling of growth plate maturation and bone formation in mice lacking both Schnurri-2 and Schnurri-3. *Proc. Natl Acad. Sci. USA* **107,** 8254–8258 (2010).
20. Ebert, P. J. R., Ehrlich, L. I. R. & Davis, M. M. Low ligand requirement for deletion and lack of synapses in positive selection enforce the gauntlet of thymic T cell maturation. *Immunity* **29,** 734–745 (2008).
21. Bousso, P., Bhakta, N. R., Lewis, R. S. & Robey, E. Dynamics of thymocyte–stromal cell interactions visualized by two-photon microscopy. *Science* **296,** 1876–1880 (2002).
22. Naeher, D. *et al.* A constant affinity threshold for T cell tolerance. *J. Exp. Med.* **204,** 2553–2559 (2007).
23. Cainan, B. J., Szychowski, S., Chan, F. K., Cado, D. & Winoto, A. A role for the orphan steroid receptor Nur77 in apoptosis accompanying antigen-induced negative selection. *Immunity* **3,** 273–282 (1995).
24. Hübner, A., Barrett, T., Flavell, R. A. & Davis, R. J. Multisite phosphorylation regulates Bim stability and apoptotic activity. *Mol. Cell* **30,** 415–425 (2008).
25. Willis, S. N. & Adams, J. M. Life in the balance: how BH3-only proteins induce apoptosis. *Curr. Opin. Cell Biol.* **17,** 617–625 (2005).
26. Glimcher, L. H. *et al.* I region-restricted antigen presentation by B cell–B lymphoma hybridomas. *Nature* **298,** 283–284 (1982).
27. Lucas, B. & Germain, R. N. Opening a window on thymic positive selection: developmental changes in the influence of cosignaling by integrins and CD28 on selection events induced by TCR engagement. *J. Immunol.* **165,** 1889–1895 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank R. Davis for providing Bim phospho-Thr-112 specific antibody; R. Germain for providing DAP3 and P139 cell lines; M. Winslow and E. Gallo for critical reading of the manuscript and valuable comments. T.L.S. is a Fellow of the Leukemia and Lymphoma Society and V.L. was a Fellow of the Irvington Institute. Supported by the National Institutes of Health grant AI29673 (L.H.G.) and K99AR055668 (D.C.J.).

Author Contributions T.L.S. designed the study and performed all experiments; V.L. and D.C.J. contributed to discussions and provided technical assistance; A.J.L. performed preliminary experiments; T.T. and S.I. provided reagents; T.L.S. and L.H.G. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Readers are welcome to comment on the online version of this article at www.nature.com/nature. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to L.H.G. (lglimche@hsph.harvard.edu).

©2011 Macmillan Publishers Limited. All rights reserved

METHODS

Mice. Mice were housed in the pathogen-free facility at the Harvard School of Public Health in accordance with guidelines from the Center for Animal Resources and Comparative Medicine at Harvard Medical School, and were used between 4 and 12 weeks of age unless otherwise noted. *Bim*$^{-/-}$ (ref. 17), DO11 (ref. 28), AND (ref. 29), HY (ref. 18), MHC I/II$^{-/-}$ (ref. 30), *Shn2*$^{-/-}$ (ref. 11) and *Shn3*$^{-/-}$ (ref. 9) strains have been previously described. Bone marrow chimaeras were generated as follows: BALB/c or MHC I/II$^{-/-}$ mice were irradiated with 2 doses of 550 rads and transplanted with 5×10^6 total bone marrow cells. Chimeras were analysed 6–8 weeks after transfer.

Real-time PCR. RNA was isolated using Trizol (Invitrogen) and DNase treated. A High Capacity cDNA Reverse Transcription kit (Applied Biosystems) was used to synthesize cDNA. SYBR Green Technology and a Stratagene Mx3005P thermocycler were used for real-time PCR. Values were normalized to β-actin. Sequences of primers: Shn2, 5′-TGAGCAGAGCACAGACACG-3′ and 5′-GGGCTCACTTTGTCAGAAGC-3′; β-actin, 5′-GCTCTGGCTCCTAGCACCAT-3′ and 5′-GCCACCGATCCACACAGAGT-3′; Bim, 5′-CGACAGTCTCAGGAGGAACC-3′ and 5′-CATTTGCAAACACCCTCCTT-3′; Nur77, 5′-TCCTCATCACTGATCGACACG-3′ and 5′-AGCTCTTCCACCCGACGAG-3′.

Biochemical analysis. Western blotting used standard methods and antibodies to: Bim phospho-Thr-112 (gift from R. Davis)[24], Bim phospho-Ser-65 (Cell Signaling), phospho-ERK (197G2, Cell Signaling), Erk (Millipore), NFATc1 (7A6, BD Pharmingen), phospho-Tyr (4G10, Upstate), Mcl-1 (Biolegend), Nur77 (12.14, BD Pharmingen), Bim (Stressgen), and SP1, HSP90, Bcl-$_{\text{XL}}$ and Bcl2 (Santa Cruz). For NFATC1 localization, nuclear and cytoplasmic purification was performed using the NE-PER kit (Pierce).

28. Murphy, K. M., Heimberger, A. B. & Loh, D. Y. Induction by antigen of intrathymic apoptosis of CD4$^+$CD8$^+$TCR$^{\text{lo}}$ thymocytes in vivo. *Science* **250,** 1720–1723 (1990).
29. Kaye, J. *et al.* Selective development of CD4$^+$ T cells in transgenic mice expressing a class II MHC-restricted antigen receptor. *Nature* **341,** 746–749 (1989).
30. Grusby, M. J. *et al.* Mice lacking major histocompatibility complex class I and class II molecules. *Proc. Natl Acad. Sci. USA* **90,** 3913–3917 (1993).

©2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09851

Hedgehog/Wnt feedback supports regenerative proliferation of epithelial stem cells in bladder

Kunyoo Shin[1], John Lee[1], Nini Guo[2], James Kim[1], Agnes Lim[1], Lishu Qu[1], Indira U. Mysorekar[3] & Philip A. Beachy[1]

Epithelial integrity in metazoan organs is maintained through the regulated proliferation and differentiation of organ-specific stem and progenitor cells. Although the epithelia of organs such as the intestine regenerate constantly and thus remain continuously proliferative[1], other organs, such as the mammalian urinary bladder, shift from near-quiescence to a highly proliferative state in response to epithelial injury[2–4]. The cellular and molecular mechanisms underlying this injury-induced mode of regenerative response are poorly defined. Here we show in mice that the proliferative response to bacterial infection or chemical injury within the bladder is regulated by signal feedback between basal cells of the urothelium and the stromal cells that underlie them. We demonstrate that these basal cells include stem cells capable of regenerating all cell types within the urothelium, and are marked by expression of the secreted protein signal Sonic hedgehog (Shh). On injury, Shh expression in these basal cells increases and elicits increased stromal expression of Wnt protein signals, which in turn stimulate the proliferation of both urothelial and stromal cells. The heightened activity of this signal feedback circuit and the associated increase in cell proliferation appear to be required for restoration of urothelial function and, in the case of bacterial injury, may help clear and prevent further spread of infection. Our findings provide a conceptual framework for injury-induced epithelial regeneration in endodermal organs, and may provide a basis for understanding the roles of signalling pathways in cancer growth and metastasis.

The multi-layered bladder epithelium consists of a lumenal layer of fully differentiated, usually binucleate umbrella cells[5], which overlie intermediate cells with limited proliferative potential, and long-term label-retaining basal cells able to produce large colonies on culture *in vitro*[6]. These urothelial layers are separated by a basement membrane from the lamina propria, a thin layer of fibroblast-like stromal cells, and submucosal, smooth muscle and serous layers. Infections of the urinary tract, occurring in 10% of women annually[7], are a common cause of injury to the bladder, and can be modelled in mice by infection with uropathogenic bacteria isolated from patients[2,3,8].

To establish baseline parameters of the bladder regenerative response, we induced injury in female mice by transurethral instillation of UTI89 (Supplementary Figs 1, 2a, 3a, b), a uropathogenic strain of *Escherichia coli*, or of the injurious compound protamine sulphate (PS). We found that expression of the proliferative marker Ki67 increased from near zero to 72% of epithelial and 28% of stromal cells within 24 h of UTI89 infection (Fig. 1a, b). The number of urothelial cell layers and total cells expressing the basal cell marker cytokeratin 5 (Ck5, also known as Krt5) was markedly expanded (Supplementary Fig. 3c, d), suggesting that injury-induced proliferation occurs primarily in basal urothelial cells. Instillation of increasing PS concentrations also induced Ki67 expression, reaching a plateau of ~35% at 20 mg ml^{-1} PS and above (Fig. 1c and Supplementary Fig. 4); Ck5-positive cell number increased more modestly with PS injury. Interestingly, stromal cells showed no increased proliferation, even at PS concentrations that saturate the epithelial response (Fig. 1c).

Shh is expressed in and has a role in development of urogenital sinus derivatives, including bladder and prostate[9–11]. In the adult urothelium, we found that *Shh* is expressed primarily in Ck5-positive basal cells, as indicated by immunostaining and GFP expression in a *Shh-GFP* BAC transgenic strain (Fig. 1d and Supplementary Fig. 5a). Hedgehog (Hh) pathway activity, indicated by a reporter *Gli1-LacZ* strain[12] (Supplementary Fig. 6a, b), is restricted to stromal, submucosal and muscle layers, outside the urothelium (Fig. 1e), and depends on the Shh signal,

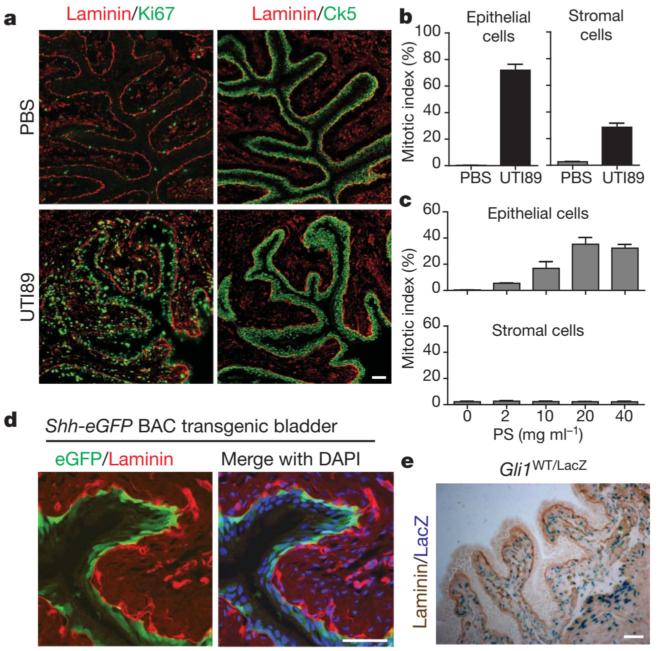


Figure 1 | Injury-induced proliferation and Hedgehog signalling in the bladder. **a**, UTI89 instillation induces proliferation of basal epithelial and stromal cells of the bladder. Ki67, Ck5 and laminin immunostaining highlight proliferation, basal epithelial cells and the basement membrane, respectively, in bladders 24 h after instillation of UTI89. Adjacent sections were 10 μm apart. **b**, Quantification of epithelial and stromal cell proliferation in response to bacterial injury. Ki67-positive cells are shown as a per cent of total 4′,6-diamidino-2-phenylindole (DAPI)-staining nuclei. **c**, Quantification of epithelial and stromal cell proliferation in response to chemical injury. Ki67-positive cells are shown as a per cent of total DAPI-staining nuclei 24 h after instillation of the indicated concentrations of PS. Note the absence of a proliferative response in the stroma. For panels **b** and **c**, data are from 3 bladders, 2 sections each, and are shown as mean ± s.e.m.; numerical data are in Supplementary Tables 1 and 2, respectively. **d**, Expression of eGFP in basal epithelial cells from a *Shh-eGFP* BAC transgenic mouse. **e**, *Gli1-LacZ* expression in the stromal compartment. Bladder sections from *Gli1*$^{\mathrm{LacZ/WT}}$ mice were co-stained with X-gal and anti-laminin. Scale bars in panels **a**, **d** and **e** represent 50 μm.

[1] Department of Developmental Biology, Institute for Stem Cell Biology and Regenerative Medicine, Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA. [2] Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [3] Department of Obstetrics and Gynecology, Washington University School of Medicine, St. Louis, Missouri 63110, USA.

©2011 Macmillan Publishers Limited. All rights reserved

as indicated by reduced *Gli1* and *Ptc* (also known as *Ptch1*) expression on treatment with a Shh-blocking antibody (Supplementary Fig. 5b, c).

Shh and *Gli1* mRNA levels both increased in response to injury (Supplementary Fig. 5d and Supplementary Table 3), and Shh protein expression extended to multiple layers of Ki67-positive epithelial cells, including basal cells and Ck5-positive basal-like cells that result from injury-induced proliferation (Supplementary Fig. 5e). Shh response was augmented by injury, requiring only four instead of twenty hours of X-gal staining for detection in *Gli1-LacZ* reporter mice, but nevertheless remained confined to the stromal compartment (Supplementary Fig. 5f).

Basal cells have been suggested to function as stem cells in many epithelia including bladder and prostate[6,13]; we marked *Shh*-expressing cells *in vivo* using a CreER tamoxifen-dependent site-specific recombinase expressed under the control of the *Shh* promoter (*Shh*$^{\mathrm{CreER}}$)[14] in combination with *R26*$^{\mathrm{mTmG}}$, a Cre-sensitive bi-fluorescent reporter[15]. In *Shh*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$ mice, the membrane-associated tomato fluorescent protein (mT) is expressed until tamoxifen (TM) injection (Supplementary Figs 2b, 7a), after which membrane-associated GFP (mG) marks Shh-expressing Ck5-positive cells in basal urothelium (Supplementary Fig. 8a). With three rounds of bacterial injury and recovery after TM injection (Supplementary Fig. 2b), mG labelled all or most urothelial cells including Ck5-positive basal cells, intermediate cells, and Ck5-negative luminal umbrella cells marked by expression of uroplakin 3 (ref. 16; Fig. 2a and Supplementary Figs 7c, 8a), indicating multipotency of Shh-expressing cells. With reduced TM treatment and a single round of bacterial injury, we observed less extensive marking in isolated, vertically coherent patches that included basal, intermediate and luminal cells (data not shown); overall regeneration thus appears to result from the combined activation of many local urothelial units.

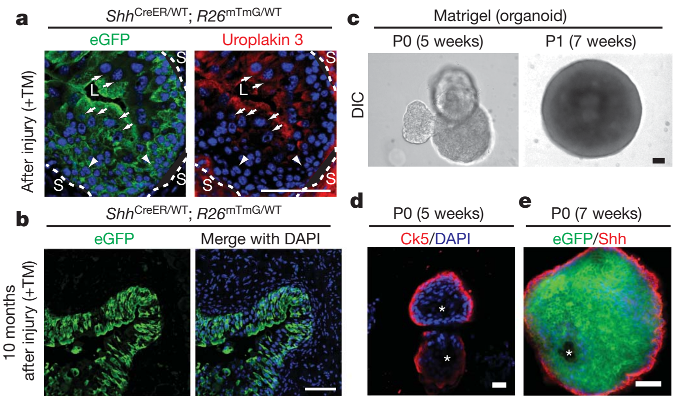


Figure 2 | Shh-expressing basal cells repopulate the urothelium and form organoids *in vitro*. **a**, eGFP-marked Shh-expressing cells generate luminal cells positive for uroplakin 3. *Shh*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$ mice were treated with TM and subjected to three rounds of injury. Uroplakin-3-positive and negative cells are denoted by white arrows and arrowheads, respectively. Dotted lines demarcate urothelium and stroma (L, lumen; S, stroma). **b**, Long-term regenerative capacity of eGFP-marked Shh-expressing cells. After TM injection and seven rounds of injury over 10 months, mG expression in bladder from *Shh*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$ mice marks most urothelial cells. See Supplementary Fig. 2b for experimental schemes. **c**, Extended culture of Shh-expressing cells in Matrigel. Single eGFP-positive bladder cells isolated from TM-injected *Shh*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$ mice and cultured in Matrigel for 5 weeks formed organoids. Dissociated cells from primary culture (P0) organoids also generated new organoids on subsequent passage (P1) in Matrigel culture. DIC, differential interference contrast. **d**, Confocal analysis of a bladder organoid. The organoid has multiple layers of epithelial cells with Ck5-expressing cells in the outer layer that contacts the extracellular matrix, and inner cells that line a luminal space and do not express Ck5. **e**, A section through the wall of an organoid grown in Matrigel and immunostained for Shh. Note eGFP expression in all cells, indicative of Shh expression in the cell initially cultured, but loss of Shh immunostaining from cells that are not in the outer layer. Scale bars represent 50 μm; asterisks denote organoid lumen.

We found similar labelling in the majority of the urothelium in mice carrying marked Shh-expressing cells and subjected to seven cycles of bacterial infection and recovery over a period of 10 months (Fig. 2b and Supplementary Fig. 8b). This persistence through lengthy and repeated periods of intense proliferation suggests that Shh-expressing cells have a capacity for self-renewal. Similarly marked cells traced through a 10-month period without injury produced less extensive labelling that nevertheless included Ck5-positive and Ck5-negative cells (Supplementary Fig. 8c, d), indicating that Shh-expressing cells also participate in regular homeostatic turnover in the absence of injury.

Similar experiments using *Gli1-CreER*[17] to mark Shh-responsive stromal cells revealed that mG expression marks most cells of the lamina propria, and no urothelial cells (Supplementary Figs 7b, 9a, b). We also timed TM administration such that cells were marked during the proliferative response to injury, and noted no injury-induced plasticity with regard to segregation of Shh and Gli1 expression within epithelial and stromal compartments, respectively (Supplementary Fig. 10a–c).

Enhanced GFP (eGFP)-positive cells isolated by fluorescence-activated cell sorting (FACS) from the bladders of TM-injected *Shh*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$ mice (WT, wild type; Supplementary Figs 2c, 11a, b) formed spheres within 2 weeks of culture in suspension or in Matrigel (Supplementary Fig. 11a–d). Approximately 5–6% of isolated cells formed primary spheres in culture, with most of the remaining cells dying rapidly, probably as a result of stress resulting from the isolation procedure; about 30–40% of cells within primary spheres were able to form secondary spheres in subsequent cultures (Supplementary Fig. 11e, f). After 5–7 weeks of culture in Matrigel, single Shh-expressing cells formed cyst-like organoids 700–1,000 μm in diameter that resemble the bladder in containing multiple layers of epithelial cells with Ck5- and Shh-expressing cells in the outer layer that contacts the extracellular matrix, and inner cells that line a luminal space and express neither Ck5 nor Shh (Fig. 2c–e, Supplementary Fig. 11g, i and Supplementary Movie 1). Single cells from these organoids were capable of self-renewing by generating new organoids in subsequent cultures (Fig. 2c, Supplementary Fig. 11h, j and Supplementary Movie 2). Our *in vivo* and *in vitro* evidence thus indicates that Shh-expressing basal urothelial cells include multipotent stem cells that are capable of self-renewal and differentiation.

The *Gli1* member of the Gli family of transcriptional effectors that mediate transcriptional response to Hh signalling (Supplementary Fig. 6a, b), although not essential for viability or fertility[12], can contribute significantly to pathway activity. For example, the incidence of medulloblastoma in $Ptc^{+/-}$ mice is reduced ~10-fold on deletion of the *Gli1* gene[18]. We found that epithelial proliferation induced by UTI89 instillation was nearly absent in *Gli1* mutant bladders at the 24–48 h peak of wild-type proliferation, with no additional layers of basal-like cells; a later peak in proliferation was ~1/2 of the wild-type maximum (Fig. 3a, b and Supplementary Fig. 12a). Proliferation of stromal cells was also affected in *Gli1* mutants, with a similar delay and decrease in Ki67 expression (Fig. 3a, b). Injection of a Shh-blocking antibody reduced expression of pathway targets *Gli1* and *Ptc* (Supplementary Fig. 5b, c), and correspondingly reduced proliferative responses in both epithelial and stromal compartments (Supplementary Fig. 12b–d and Supplementary Table 5). The response to chemical injury was also markedly reduced in *Gli1* mutants (Fig. 3c and Supplementary Fig. 13a, b), and these results together indicate that stromal Gli1-mediated response to the epithelial Shh signal promotes proliferative activity in response to both chemical and bacterial injury.

We tested the role of proliferation in restoring urothelial integrity by instilling fluorescein isothiocyanate (FITC)-conjugated dextran after UTI89-mediated injury in wild-type and *Gli1* mutants. At 6 h after infection, by which time umbrella cells have exfoliated[8], wild-type and mutant bladders both showed penetration of FITC-dextran into interstitial spaces of the urothelium (not shown); at 24 h, however,

©2011 Macmillan Publishers Limited. All rights reserved

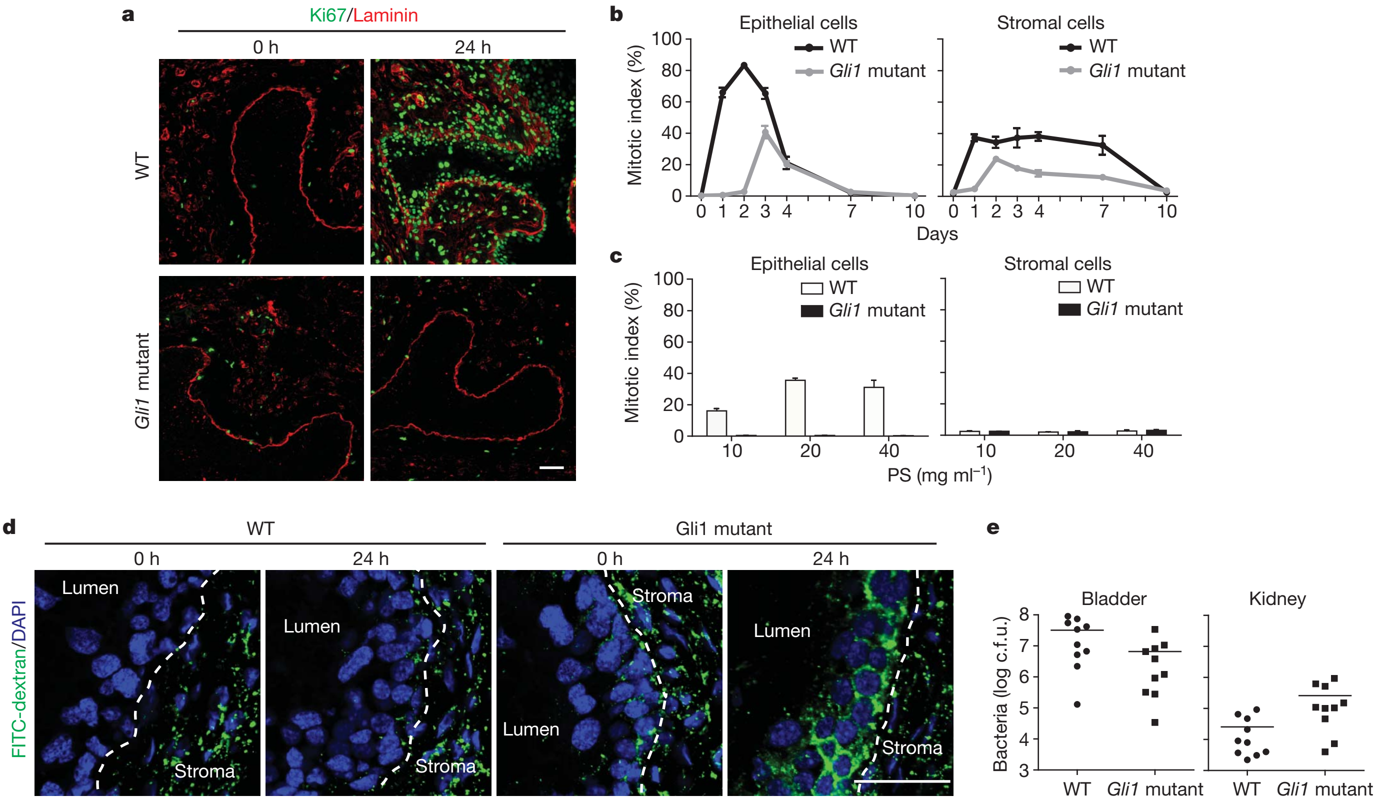


Figure 3 | Gli1 mediates injury-induced proliferation, restoration of urothelial integrity and reduction of infectious spread. a, *Gli1* loss delays and attenuates proliferative response to bacterial injury. Bladders from wild-type or homozygous *Gli1* mutant mice were analysed at the indicated times after UTI89 instillation by immunostaining for Ki67 and laminin. **b**, Quantification of *Gli1* effect on epithelial and stromal cell proliferation. Ki67-positive cells are shown as a per cent of total DAPI-staining nuclei at the indicated times after UTI89 instillation in wild-type and *Gli1* mutant bladders. **c**, *Gli1* loss blocks the proliferative response of epithelial cells to chemical injury. Ki67-positive cells are shown as a per cent of total DAPI-staining nuclei 24 h after instillation of the indicated concentrations of PS. In panels **b** and **c**, data are shown as mean ± s.e.m. from 3 bladders, 2 sections each, and numerical data are shown in Supplementary Tables 4 and 6, respectively. **d**, Paracellular permeability in injured bladders from *Gli1* mutant mice. Bladders from wild-type and *Gli1* homozygotes were instilled with UTI89 and analysed at the times indicated after infection, with FITC-dextran instillation preceding bladder collection by 1.5 h. Dotted lines demarcate the border between urothelium and stroma. Note that normal reduced levels of paracellular permeability are restored by 24 h in wild-type but not *Gli1* homozygotes. **e**, Infectious spread to kidneys is enhanced by *Gli1* loss. Bacterial titres 24 h after UTI89 instillation were lower in bladders from *Gli1* homozygotes as compared to wild-type ($6.6 \pm 3.27 \times 10^6$ versus $3.2 \pm 1.03 \times 10^7$ colony-forming units (c.f.u.); $P < 0.05$). In contrast, bacterial titres were significantly higher in kidneys from *Gli1* homozygotes as compared to wild-type ($2.59 \pm 1.0 \times 10^5$ versus $2.55 \pm 1.0 \times 10^4$; $P < 0.05$). Data are presented as mean ± s.e.m. (10 mice), and significance was calculated by an unpaired Student's *t*-test. Scale bars represent 50 μm.

wild-type but not mutant bladders had re-established exclusion of FITC-dextran from extracellular spaces (Fig. 3d).

We also found that, although bacterial titres were somewhat lower in the bladders of infected *Gli1* mutant mice (Fig. 3e), the kidneys of mutant mice contained more than tenfold higher numbers of bacteria (Fig. 3e). These findings indicate that in addition to helping restore epithelial integrity, rapid proliferation of urothelial cells during normal regeneration may help reduce the risk of bacterial spread from the bladder to the kidneys, perhaps by competing for adhesive interactions that otherwise might aid in bacterial ascent via the ureters to the kidneys[19].

The requirement for stromal Gli1 in mediating proliferative response to epithelial injury suggested the possibility of Shh/Gli1-dependent transcription of secreted signals within the stroma. From gene expression profiles (data not shown) and quantitative polymerase chain reaction with reverse transcription (RT–PCR) with RNA from injured and uninjured wild-type or *Gli1* mutant bladders, we found that *Wnt2*, *Wnt4* and *Fgf16* showed significant injury- and *Gli1*-dependent responses (Supplementary Fig. 14a, b). With RNA isolated from stromal and epithelial compartments using laser capture microdissection (LCM; Supplementary Fig. 15a, b), we found that levels of *Wnt2*, *Wnt4* and *Fgf16* transcripts increased with injury only in the stroma (Fig. 4a).

For further analysis we isolated basal, intermediate and umbrella cell layers from uninjured wild-type or *Gli1* mutant bladders (Supplementary Fig. 15c, d); as umbrella cells were absent from regenerating bladders owing to bacterially induced exfoliation, we isolated a single lumenal layer of intermediate cells, a single layer of Ck5-positive basal cells and two layers of Ck5-positive basal-like cells (basal-like1 and basal-like2) from wild type (Fig. 4b and Supplementary Fig. 15e) or, from *Gli1* homozygous mutants, intermediate cells and a single layer of basal cells (Supplementary Fig. 15f). With RNA from these microdissected cell layers we found that transcription of *Axin2*, a universal indicator of Wnt signal response[20,21], increased markedly in stromal cells and in basal and basal-like1 cells, to a lesser extent in the second basal-like2 layer, and not at all in the intermediate cell layer (Fig. 4c). No increase in *Axin2* levels could be seen in stromal or urothelial layers of injured, *Gli1*-mutant bladders (Fig. 4c). Increased Wnt response thus occurs in stromal cells and in urothelial cell layers in closest proximity to the stromal source of Wnt signals.

Shh expression also increased in basal and basal-like cells (Fig. 4d and Supplementary Fig. 16), indicating stromal Gli1-mediated positive feedback on epithelial *Shh* expression. Notably, however, *Shh* expression in basal cells also increased somewhat in the *Gli1* mutant, indicating an injury-induced effect on *Shh* expression that does not require Gli1-mediated feedback from the stroma.

We tested pharmacological modulators of Wnt signalling, and found that indomethacin treatment reduced *Axin2* transcripts twofold in the bladder (Supplementary Fig. 17a), indicating a reduction in Wnt

©2011 Macmillan Publishers Limited. All rights reserved

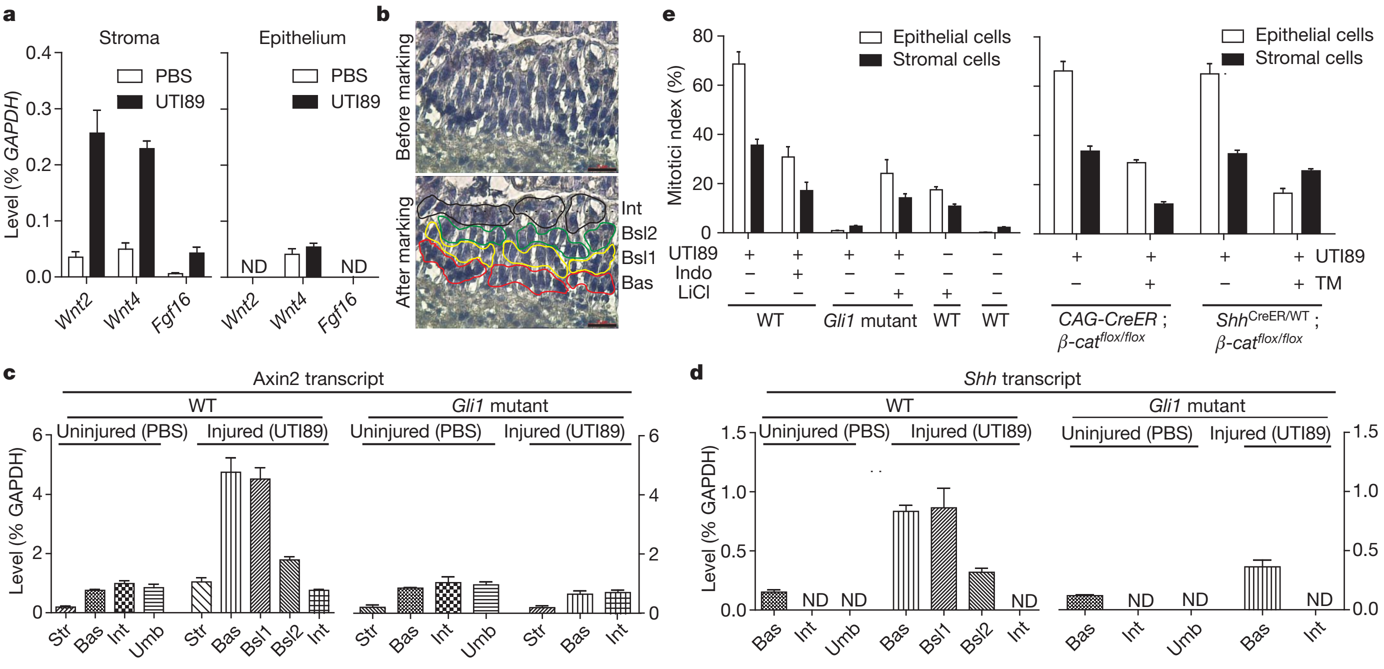


Figure 4 | Hedgehog-induced expression of stromal Wnt signals mediates urothelial and stromal proliferation. **a**, *Wnt2*, *Wnt4* and *Fgf16* expression in microdissected epithelium or stroma. Stromal expression of *Wnt2*, *Wnt4* and *Fgf16* increased significantly on injury. Although *Wnt4* RNA was detected in the epithelium, this expression did not increase on injury. ND, not detected. **b**, Laser capture microdissection of urothelial cell layers in regenerating bladder. Red, yellow, green and black lines illustrate selection and outlining of cells in basal (Bas), basal-like 1 (Bsl1), basal-like 2 (Bsl2), and intermediate (Int) layers, respectively, before microdissection. **c**, **d**, *Axin2* and *Shh* expression in microdissected cell layers from wild-type and *Gli1* mutant bladders 24 h after instillation of UTI89 or PBS. **c**, Expression of *Axin2* increased 5.5 fold in stroma ($P < 0.05$), and 6.3 fold in basal ($P < 0.05$), 5.9 fold in basal-like 1 ($P < 0.05$), and 2.3 fold in basal-like 2 ($P < 0.01$) as compared to uninjured basal cells. Expression of *Axin2* did not change significantly in basal and intermediate cells of injured *Gli1* homozygous mutants as compared to basal and intermediate cells of uninjured *Gli1* mutant bladder. **d**, Expression of *Shh* in cells increased 5.5 fold in basal ($P < 0.01$), 5.7 fold in basal-like 1 ($P < 0.05$), and 2.1 fold in basal-like 2 ($P < 0.05$) as compared to uninjured basal cells. Expression of *Shh* in *Gli1* homozygous mutants increased 3 fold ($P < 0.05$) in injured as compared to uninjured basal epithelial cells. Data are presented as mean ± s.e.m., and significance was calculated by a paired Student's *t*-test. Str, stroma; Umb, umbrella cells. **e**, Modulation of Wnt pathway activity in regenerative response to bacterial injury. Left, pharmacological reduction (indomethacin (Indo)) or augmentation (LiCl) of the Wnt signal response respectively decreases or increases bladder proliferation, with or without bacterial injury, as shown, in mice of the indicated genotype (see also Supplementary Fig. 17b, c, d). Right, tamoxifen-induced inactivation of β-catenin decreases proliferation in epithelium and stroma (*CAG-CreER*) or in epithelium (Shh^{CreER}) . Data are presented as mean ± s.e.m. from 3 bladders, 2 sections each (see also Supplementary Fig. 18e, f).

response[20], and correspondingly suppressed UTI89-induced proliferation in the epithelium and stroma (Fig. 4e and Supplementary Fig. 17b). We also found that LiCl[22] increased *Axin2* transcripts (Supplementary Fig. 17a), and correspondingly induced proliferation in the epithelium and stroma of uninjured wild-type mice or substantially rescued the proliferative response of *Gli1* mutant mice (Fig. 4e and Supplementary Fig. 17c, d). We also observed a marked enhancement of the proliferative response to instillation of 2 mg ml^{-1} PS, from ~5% to 30% in the epithelium and from 0 to ~15% in the stroma (Supplementary Fig. 18a, b), in mice heterozygous for the Apc^{min} mutation, which show constitutive Wnt pathway activity due to reduced dosage of a negative Wnt response regulator[23] (Supplementary Fig. 18c).

We inactivated the Wnt response with a homozygous conditional allele of the essential Wnt pathway component β-catenin (also known as Ctnnb1), and found that ablation in TM-injected mice by the ubiquitously expressed *CAG-CreER* reduced *Axin2* expression in injured bladder (Supplementary Fig. 18d) and correspondingly reduced injury-induced proliferation in both epithelial and stromal compartments (Fig. 4e and Supplementary Fig. 18e). In contrast, the proliferation defect produced in $Shh^{\mathrm{CreER/WT}}$; $\beta\text{-}catenin^{flox/flox}$ mice was restricted to basal epithelium (Fig. 4e and Supplementary Fig. 18f). Our pharmacological and genetic data indicate a role for Wnt pathway activity within basal epithelial and stromal cells in the activation of proliferative responses, with the potential role and importance of Fgf or other Hh-induced stromal signals remaining to be explored.

Our findings, summarized schematically in Supplementary Fig. 19, reveal an essential contribution by Hh and Wnt signals acting across the epithelial–stromal boundary during bladder regeneration, and are reminiscent of the well-studied *Drosophila* embryonic segment, in which Hh and Wingless signalling across the parasegment boundary are essential for segmental patterning during development. Interestingly, however, Hh/Wnt feedback signalling in the *Drosophila* segment operates to specify future pattern within an undifferentiated epithelium, whereas in the bladder it functions to maintain differentiated structures in a mature organ.

Surprisingly, despite its dispensability for normal development, *Gli1* contributes significantly to bladder regeneration, thus providing a useful tool to demonstrate a role of regenerative proliferation not only in restoring urothelial integrity but also in preventing bacterial spread to the kidneys. Further studies will be required to determine whether Gli1 also contributes to the regeneration of other organs in which Hh signalling has a role[11,24]; such flexibility in Gli1 transcriptional output would be consistent with the apparent flexibility of Gli1 in mediating qualitatively distinct responses to bacterial and chemical bladder injury, which differ markedly in the presence or absence of stromal cell proliferation.

As many as 10% of women experience infections of the urinary tract in a year[7,25], including cystitis and pyelonephritis, and ~26% of urinary tract infections recur within six months[26]. Recent work has demonstrated that intracellular reservoirs of bacteria can form in umbrella cells or in transitional cells[3], and exfoliation of infected cells harbouring intracellular bacterial reservoirs induced by PS or other agents has been suggested as potentially beneficial in clearing these infections. Our current data suggest that it may also be worth exploring the effect of

©2011 Macmillan Publishers Limited. All rights reserved

more directly activating the Wnt signalling pathway by treating with LiCl; on the other hand use of drugs such as indomethacin, which inhibit the regenerative response in the urothelium, may be contraindicated.

Another clinical arena in which our findings may be relevant is the growth and dissemination of cancers in which Hh ligand production in primary cells of the tumour triggers pathway activity in tumour stroma, which then expands and supports the growth of primary cells within the tumour[27,28]. This tumour–stromal interaction might now be viewed as the growth-enhancing activation of a feedback circuit normally triggered by injury. Tumour–stromal interactions are also critical in metastasis, and the preferential colonization of particular tissues by individual tumour types might be determined by the competence of target tissues to respond to eliciting signals from the tumour cells, similar to the epithelial–stromal interaction noted here in bladder regeneration. In this connection it is interesting to note that tumours of the prostate, which like the bladder is a derivative of the urogenital sinus and also requires Hh pathway activity for its regeneration[11], metastasize most commonly to bone, lung and liver[29], a preference very similar to that of bladder tumours, which metastasize most commonly to liver, lung and bone[30]. Further work will be required to identify regenerative signals in other organs and to determine how important their activities may be in cancer growth and metastasis.

METHODS SUMMARY

For bladder injury with chemical agents or bacteria, transurethral instillation was performed as described[2]. For *in vivo* lineage tracing, transgenes expressing tamoxifen-activated Cre combined with the *R26*$^{\mathrm{mTmG}}$ bi-fluorescent reporter were used to heritably mark cells and their progeny. Transgenes expressing tamoxifen-activated Cre also were used to selectively inactivate a conditional allele of β-catenin. For bladder sphere and organoid culture, GFP-marked Shh-expressing basal cells from TM-treated *Shh*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$ mice were FACS-sorted and cultured in suspension or in Matrigel. Quantitative RT–PCR was used for measurement of RNA levels in samples isolated from the entire bladder, or in material isolated by laser capture microdissection from the stromal or the urothelial compartments, or from individual cell layers within the urothelium.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 September 2010; accepted 17 January 2011.
Published online 9 March 2011.

1. van der Flier, L. G. & Clevers, H. Stem cells, self-renewal, and differentiation in the intestinal epithelium. *Annu. Rev. Physiol.* **71,** 241–260 (2009).
2. Hung, C. S., Dodson, K. W. & Hultgren, S. J. A murine model of urinary tract infection. *Nature Protocols* **4,** 1230–1243 (2009).
3. Mysorekar, I. U. & Hultgren, S. J. Mechanisms of uropathogenic *Escherichia coli* persistence and eradication from the urinary tract. *Proc. Natl Acad. Sci. USA* **103,** 14170–14175 (2006).
4. Mysorekar, I. U., Isaacson-Schmid, M., Walker, J. N., Mills, J. C. & Hultgren, S. J. Bone morphogenetic protein 4 signaling regulates epithelial renewal in the urinary tract in response to uropathogenic infection. *Cell Host Microbe* **5,** 463–475 (2009).
5. Hicks, R. M. The mammalian urinary bladder: an accommodating organ. *Biol. Rev. Camb. Philos. Soc.* **50,** 215–246 (1975).
6. Kurzrock, E. A., Lieu, D. K., Degraffenried, L. A., Chan, C. W. & Isseroff, R. R. Label-retaining cells of the bladder: candidate urothelial stem cells. *Am. J. Physiol. Renal Physiol.* **294,** F1415–F1421 (2008).
7. Hooton, T. M. & Stamm, W. E. Diagnosis and treatment of uncomplicated urinary tract infection. *Infect. Dis. Clin. North Am.* **11,** 551–581 (1997).
8. Klumpp, D. J. *et al.* Uropathogenic *Escherichia coli* induces extrinsic and intrinsic cascades to initiate urothelial apoptosis. *Infect. Immun.* **74,** 5106–5113 (2006).
9. Podlasek, C. A., Barnett, D. H., Clemens, J. Q., Bak, P. M. & Bushman, W. Prostate development requires Sonic hedgehog expressed by the urogenital sinus epithelium. *Dev. Biol.* **209,** 28–39 (1999).
10. Haraguchi, R. *et al.* Molecular analysis of coordinated bladder and urogenital organ formation by Hedgehog signaling. *Development* **134,** 525–533 (2007).
11. Karhadkar, S. S. *et al.* Hedgehog signalling in prostate regeneration, neoplasia and metastasis. *Nature* **431,** 707–712 (2004).
12. Bai, C. B., Auerbach, W., Lee, J. S., Stephen, D. & Joyner, A. L. Gli2, but not Gli1, is required for initial Shh signaling and ectopic activation of the Shh pathway. *Development* **129,** 4753–4761 (2002).
13. Lawson, D. A., Xin, L., Lukacs, R. U., Cheng, D. & Witte, O. N. Isolation and functional characterization of murine prostate stem cells. *Proc. Natl Acad. Sci. USA* **104,** 181–186 (2007).
14. Harfe, B. D. *et al.* Evidence for an expansion-based temporal Shh gradient in specifying vertebrate digit identities. *Cell* **118,** 517–528 (2004).
15. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45,** 593–605 (2007).
16. Wu, X. R. & Sun, T. T. Molecular cloning of a 47 kDa tissue-specific and differentiation-dependent urothelial cell surface glycoprotein. *J. Cell Sci.* **106,** 31–43 (1993).
17. Ahn, S. & Joyner, A. L. Dynamic changes in the response of cells to positive Hedgehog signaling during mouse limb patterning. *Cell* **118,** 505–516 (2004).
18. Kimura, H., Stephen, D., Joyner, A. & Curran, T. Gli1 is important for medulloblastoma formation in *Ptc1*$^{+/-}$ mice. *Oncogene* **24,** 4026–4036 (2005).
19. Hagberg, L. *et al.* Ascending, unobstructed urinary tract infection in mice caused by pyelonephritogenic *Escherichia coli* of human origin. *Infect. Immun.* **40,** 273–283 (1983).
20. Goessling, W. *et al.* Genetic interaction of PGE2 and Wnt signaling regulates developmental specification of stem cells and regeneration. *Cell* **136,** 1136–1147 (2009).
21. Lustig, B. *et al.* Negative feedback loop of Wnt signaling through upregulation of conductin/axin2 in colorectal and liver tumors. *Mol. Cell. Biol.* **22,** 1184–1193 (2002).
22. Klein, P. S. & Melton, D. A. A molecular mechanism for the effect of lithium on development. *Proc. Natl Acad. Sci. USA* **93,** 8455–8459 (1996).
23. Su, L. K. *et al.* Multiple intestinal neoplasia caused by a mutation in the murine homolog of the APC gene. *Science* **256,** 668–670 (1992).
24. Fendrich, V. *et al.* Hedgehog signaling is required for effective regeneration of exocrine pancreas. *Gastroenterology* **135,** 621–631 (2008).
25. Nicolle, L. E. Uncomplicated urinary tract infection in adults including uncomplicated pyelonephritis. *Urol. Clin. North Am.* **35,** 1–12 (2008).
26. Foxman, B. Recurring urinary tract infection: incidence and risk factors. *Am. J. Public Health* **80,** 331–333 (1990).
27. Yauch, R. L. *et al.* A paracrine requirement for hedgehog signalling in cancer. *Nature* **455,** 406–410 (2008).
28. Tian, H. *et al.* Hedgehog signaling is restricted to the stromal compartment during pancreatic carcinogenesis. *Proc. Natl Acad. Sci. USA* **106,** 4254–4259 (2009).
29. Bubendorf, L. *et al.* Metastatic patterns of prostate cancer: an autopsy study of 1,589 patients. *Hum. Pathol.* **31,** 578–583 (2000).
30. Wallmeroth, A. *et al.* Patterns of metastasis in muscle-invasive bladder cancer (pT2–4): an autopsy study on 367 patients. *Urol. Int.* **62,** 69–75 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Oro and J. Brooks for their critical reading of the manuscript, and the Stanford Center for Digestive Diseases for help with laser capture microdissection. This research was supported in part by grants from the Department of Defense and from the National Institutes of Health (P.A.B.) and a Pathway to Independence Award (K99/R00) to I.U.M. P.A.B. is an investigator of the Howard Hughes Medical Institute.

Author Contributions K.S. and P.A.B. conceived ideas and experimental design. K.S. performed the experiments. N.G. aided in immunohistochemical analysis, J.L. and J.K. helped with mouse strains, A.L. assisted with *in vitro* cell culture studies, L.Q. performed the genotyping of experimental mice, and I.U.M. helped analyse data. K.S. and P.A.B wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to P.A.B. (pbeachy@stanford.edu) or K.S. (kunyoos@stanford.edu).

©2011 Macmillan Publishers Limited. All rights reserved

METHODS

Mice. *Gli1*$^{\mathrm{LacZ/WT}}$ heterozygotes[12] from our colony were interbred to generate *Gli1*$^{\mathrm{LacZ/LacZ}}$ homozygotes. Heterozygous and wild-type littermates were used as controls. Female mice between 6 and 10 weeks of age were used for all injury experiments. *Shh-GFP* BAC strain was obtained from the GENSAT project at Rockefeller University. For lineage tracing experiments, *Shh*$^{\mathrm{CreER/WT}}$, or *Gli1*$^{\mathrm{CreER/WT}}$ mice[14,17] were crossed with *R26*$^{\mathrm{mTmG/mTmG}}$ strain[15] to obtain *Shh*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$, or *Gli1*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$. *Shh*$^{\mathrm{CreER/WT}}$ or *CAG-CreER* were crossed with *β-catenin*$^{flox/flox}$ mice to obtain *Shh*$^{\mathrm{CreER/WT}}$; β-catenin$^{flox/WT}$ or *CAG-CreER*; β-catenin$^{flox/WT}$. Resulting mice were crossed with β-catenin$^{flox/flox}$ mice to obtain *Shh*$^{\mathrm{CreER/WT}}$; β-catenin$^{flox/flox}$ or *CAG-CreER*; β-catenin$^{flox/flox}$. All mouse strains except as otherwise indicated were obtained from Jackson Laboratories. All bladder instillation procedures were performed under isoflurane anaesthesia, which was administered in a fume hood with a standard vaporizer (J. B. Baulch and Associates). All procedures were performed under a protocol approved by the Administrative Panel on Laboratory Animal Care at Stanford University.

Bacterial and chemical injury. For bacterial injury, a uropathogenic *E. coli* (UPEC) strain, UTI89, was grown for 16 h in a static culture, and inoculated via transurethral instillation of anaesthetized female mice at a concentration of 10^7 c.f.u. in 50 µl as previously described[2]. Animals were maintained after infection for the indicated period of time or urine was collected after 4 h to confirm infection. For chemical injury, 50 µl of PS (Sigma) solution in PBS at 2 mg ml^{-1}, 10 mg ml^{-1}, 20 ml ml^{-1} or 40 mg ml^{-1} was delivered transurethrally as indicated. Bladders were collected and analysed at distinct time-points after instillation.

Lineage tracing studies. For fate mapping of Shh-expressing cells before bacterial injury, *Shh*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$ or *Gli1*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$ mouse strains were injected intraperitoneally with 4 mg of TM (per 30 g body weight) daily for three consecutive days. Seven days after the last TM injection, mice were subjected to three transurethral instillations of UTI89 with 10-day intervals after the first and second infections, and 15 days after the third. Mice were killed and bladders then dissected for further analysis. For labelling of *Shh*- or *Gli1*-expressing cells before injury, mice were injected with TM (4 mg per 30 g body weight) for 3 days and analysed 5 days after the last injection. For lineage tracing of Shh or Gli1 expressing cells during injury, *Shh*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$ or *Gli1*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$ mice were injected intraperitoneally with 4 mg of TM (per 30 g body weight) for 5 consecutive days, starting 2 days before infections to allow enough time for tamoxifen to be absorbed by the bladder. Six days after the last injection of TM, the entire procedure was repeated. Tissue sections were prepared 6 days after final injection of TM for further analysis. For long-term lineage tracing, *Shh*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$ mouse strains were injected intraperitoneally with 4 mg of TM (per 30 g body weight) daily for 3 consecutive days. Seven days after the last TM injection, mice were subjected to seven transurethral instillations of UTI89, twice a month for 2 months and once a month for next 3 months. Five months after the last instillation, mice were killed and bladders then dissected for further analysis. For long-term lineage tracing experiments to study homeostasis, *Shh*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$ mouse strains were injected with 4 mg of TM (per 30 g body weight) daily for 3 consecutive days, and bladders were analysed 10 months after the last TM injection.

***In vitro* culture of Shh-expressing cells.** To isolate Shh-expressing cells, *Shh*$^{\mathrm{CreER/WT}}$; *R26*$^{\mathrm{mTmG/WT}}$ mouse strains were injected intraperitoneally with 4 mg of TM (per 30 g body weight) daily for 3 consecutive days. Three days after TM injection, bladders were collected, inverted and inflated as described previously[31]. Inverted bladders were incubated in 0.25% Trypsin-EDTA containing 500 U ml^{-1} collagenase for 2 h at 37 °C. Tissues were then minced and, after the lysis of red blood cells, a single-cell suspension was obtained by 10 min of trituration, followed by filtration through 40-µm cell strainers. Cells were sorted using a FACS AriaII cytometer (BD Biosciences), and analysis of flow cytometry data was performed using FlowJo Software (Treestar). Sorted cells were cultured in Ultra Low attachment plates (Corning) for suspension culture in media containing 1:1 mixture of conditioned media and DMEM (Invitrogen) supplemented with 50 ng ml^{-1} EGF (PeproTech). Conditioned medium was prepared by growing 90% confluent V79 lung fibroblast-derived cells (ATCC) in DMEM with 5% FBS for 24 h. For Matrigel culture, a single-cell suspension was mixed with 200 µl of ice-cold Matrigel (reduced growth factors; BD Bioscience), layered onto 12-mm Transwell clear filters, and allowed to solidify at 37 °C. Pre-warmed growth medium was added above and below the gel and cells were grown until the time of analysis. For passaging organoids in Matrigel culture, organoids were released from Matrigel matrix by depolymerizing Matrigel using MatriSphere Cell Recovery Solution (BD Bioscience) according to the manufacturer's instructions. Organoids were then dissociated into single cells by incubating with 0.25% Trypsin-EDTA for 30 min, followed by 5 min of trituration. Cells were then seeded in Matrigel as described earlier.

Antibody injection. Mice were injected intraperitoneally with either anti-mouse Shh antibody or isotype control (5E1 and 6B3, respectively; Developmental Studies Hybridoma Bank) daily for 7 days. Mice received 10 mg kg^{-1} body weight of antibody for the first 3 days and 5 mg kg^{-1} body weight thereafter. Bladders were dissected on the last day of injection to isolate RNA for qRT–PCR analysis. For assays of proliferation, mice were injected with 10 mg antibody per kg body weight daily for 3 days, and 5 mg antibody per kg body weight thereafter. Mice were infected with UTI89 on the fourth day of antibody injection.

Microscopy and laser capture microdissection. All images were obtained using a Zeiss LSM 510 inverted confocal microscope and prepared for publication with Zeiss LSM 5 Image Browser software and Adobe Photoshop CS3. Three-dimensional reconstructions of confocal images were generated using Imaris software (Bitplane Scientific Software). For LCM, bladder sections were prepared using an LCM staining kit (Ambion) and a Leica LMD6000 Laser Microdissection Microscope.

Quantitative RT–PCR. For qRT–PCR, bladders were frozen in liquid nitrogen and total RNAs isolated from frozen bladder tissue using RNeasy Plus Mini (Qiagen). qRT–PCR was performed using iScript one-step RT–PCR kit with SYBR Green and the Bio-Rad iCycler (BioRad). Assays were performed on bladders from six animals, and all values normalized to the GAPDH internal control, which does not vary on injury (data not shown). For material isolated by LCM, total RNA was prepared using RNAqueous-Micro RNA isolation kit (Ambion).

Indomethacin and LiCl treatment. Mice were injected with 2.5 mg kg^{-1} of body weight with indomethacin or DMSO vehicle control every 12 h throughout the experiment[20]. UTI89 was instilled 36 h after initial indomethacin injection, and bladders were analysed on the third day, 60 h after the initial indomethacin injection. For LiCl treatment, wild-type or *Gli1*$^{\mathrm{LacZ/LacZ}}$ mice received either 200 mg kg^{-1} of body weight of LiCl in 100 µl of deionized water or 100 µl of deionized water as a vehicle control every day for 3 days by oral gavage. At the time of oral gavage, mice were also injected transurethrally with 40 mg kg^{-1} of body weight of LiCl. UTI89 infection was performed at 30 h after initial LiCl treatment, and bladders were analysed on the third day, 54 h after initial LiCl treatment.

Conditional ablation of β-catenin. *Shh*$^{\mathrm{CreER/WT}}$; β-catenin$^{flox/flox}$ or *CAG-CreER*; β-catenin$^{flox/flox}$ female mice were injected intraperitoneally with 4 mg of TM (per 30 g body weight) daily for 3 consecutive days. For *CAG-CreER*; β-catenin$^{flox/flox}$, bladder was inoculated with UTI89 the day after the last injection of TM, and bladders were collected 24 h after infection. For *Shh*$^{\mathrm{CreER/WT}}$; β-catenin$^{flox/flox}$, bladder was inoculated with UTI89 3 days after the last injection of TM, and collected 24 h after infection.

***In vivo* permeability assay.** 10 mg ml^{-1} of FITC-dextran (10000MW, Invitrogen) in a 50 ml volume of PBS[32] was injected transurethrally into the bladder lumen 1.5 h before collection of bladders at the indicated time points. Bladder sections were made and analysed for FITC.

Bacterial titration and analysis of urine. Urine from infected bladders was collected 6 h after UTI89 infection, and slides were prepared and stained for analysis using Cytospin and Hema3 staining kit (Fisher). Kidneys and bladders were dissected from wild-type littermate controls or *Gli1*$^{\mathrm{LacZ/LacZ}}$ mice 24 h after infection. Tissues were homogenized in 1 ml of PBS, and bacterial titres were determined by microtitre-plate dilution on LB plates[2].

X-gal histochemistry and immunofluorescence analysis. Bladders were dissected and embedded in OCT compound for snap freezing (Tissue-Tek). Frozen blocks were sectioned at 10-µm intervals using a Microm cryostat. For X-gal staining, frozen sections were fixed in 0.2% glutaraldehyde in PBS containing 5 mM EGTA and 2 mM $\mathrm{MgCl_2}$ for 30 min at 4 °C. After washing twice with PBS containing 2 mM $\mathrm{MgCl_2}$, sections were incubated with 1 mg ml^{-1} of X-gal solution in PBS containing 0.02% NP40, 0.01% deocycholic acid, 2 mM $\mathrm{MgCl_2}$, 5 mM EGTA, 5 mM $\mathrm{C_6FeK_3N_6}$, 5 mM $\mathrm{C_6FeK_4N_6}$ for 4 h to overnight. Stained sections were counterstained with eosin solution (Sigma). For immunostaining, frozen tissue sections were fixed in 4% of paraformaldehyde for 30 min at 4 °C. After washing three times with PBS, tissue sections were blocked in 2% goat serum in PBS containing 0.25% Trion X-100 for 1 h, incubated with the following primary antibodies diluted in blocking solution overnight at 4 °C in a humidified chamber: rat anti-Shh (R&D, 1:200); rabbit anti-Ki67 (Abcam, 1:500); rabbit anti-Ck5 (Abcam, 1:500); chicken anti-laminin (Abcam, 1:300); mouse anti-uroplakin 3 (Fitzgerald). Sections were washed three times with PBS containing 0.25% Triton X-100, incubated with DAPI and appropriate Alexa fluoro 488, 594, or 633 conjugated secondary antibodies diluted 1:1,000 in blocking solution for 2 h at 22 °C, washed again three times, and mounted on slides with Prolong Gold mounting reagent (Invitrogen). For immunostaining of organoids, Matrigel plugs were removed from the filter supports of Transwell plates, washed with PBS, and incubated with MatriSphere Cell Recovery Solution (BD Bioscience) for 20 min to partially dissolve the Matrigel. Subsequently, the gel-embedded bladder organoids were fixed in 4% paraformaldehyde for 15 min at 4 °C, followed by washing three times with PBS. Fixed organoids then were permeablized in 0.25% Triton X-100 for 30 min, followed by incubation in blocking solution of 2% goat serum in

©2011 Macmillan Publishers Limited. All rights reserved

PBS containing 0.25% Triton X-100 for 1 h. Organoids were then incubated with primary antibodies diluted in blocking solution for 2 days at 4 °C in a humidified chamber. After three washes, bladder organoids were incubated with secondary antibody overnight in 4 °C, followed by three washes and mounting on Coverwell chamber (Grace Bio-Lab) with Prolong Gold mounting reagent (Invitrogen).

Statistical analysis. Statistical analysis was performed using GraphPad Prism software v.5. All data are presented as mean ± s.e.m., and two group comparisons were done with a two-tailed Student's t-test. A value of $P < 0.05$ was taken as statistically significant.

31. Kurzrock, E. A., Lieu, D. K., deGraffenried, L. A. & Isseroff, R. R. Rat urothelium: improved techniques for serial cultivation, expansion, freezing and reconstitution onto acellular matrix. *J. Urol.* **173,** 281–285 (2005).
32. Ibla, J. C. & Khoury, J. Methods to assess tissue permeability. *Methods Mol. Biol.* **341,** 111–117 (2006).

©2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09861

An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress

Hidetaka Ito[1]*†, Hervé Gaubert[1]*, Etienne Bucher[1]*†, Marie Mirouze[1]†, Isabelle Vaillant[1]† & Jerzy Paszkowski[1]

Eukaryotic genomes consist to a significant extent of retrotransposons that are suppressed by host epigenetic mechanisms, preventing their uncontrolled propagation[1,2]. However, it is not clear how this is achieved. Here we show that in *Arabidopsis* seedlings subjected to heat stress, a *copia*-type retrotransposon named *ONSEN* (Japanese 'hot spring') not only became transcriptionally active but also synthesized extrachromosomal DNA copies. Heat-induced *ONSEN* accumulation was stimulated in mutants impaired in the biogenesis of small interfering RNAs (siRNAs); however, there was no evidence of transposition occurring in vegetative tissues. After stress, both *ONSEN* transcripts and extrachromosomal DNA gradually decayed and were no longer detected after 20–30 days. Surprisingly, a high frequency of new *ONSEN* insertions was observed in the progeny of stressed plants deficient in siRNAs. Insertion patterns revealed that this transgenerational retrotransposition occurred during flower development and before gametogenesis. Therefore in plants with compromised siRNA biogenesis, memory of stress was maintained throughout development, priming *ONSEN* to transpose during differentiation of generative organs. Retrotransposition was not observed in the progeny of wild-type plants subjected to stress or in non-stressed mutant controls, pointing to a crucial role of the siRNA pathway in restricting retrotransposition triggered by environmental stress. Finally, we found that natural and experimentally induced variants in *ONSEN* insertions confer heat responsiveness to nearby genes, and therefore mobility bursts may generate novel, stress-responsive regulatory gene networks.

In *Arabidopsis* mutants compromised in 24-nucleotide siRNA biogenesis, transposon transcripts appear but transposition has not been observed[3,4]. This is in contrast to mutants lacking chromatin-remodelling factor DDM1 or DNA methyltransferase MET1, in which transposons move during inbreeding[5–9]. It has been shown that transposon transcripts and their siRNAs accumulate in the vegetative nucleus of pollen[10]. A similar observation has been reported for the endosperm[11–13]. For pollen vegetative cells, where transposon mobility has been observed, it has been postulated that relocation of transposon siRNAs to sperm cells contributes to transposon silencing in the germ line[10]. Nevertheless, it is troubling that transposons remain immobile during inbreeding of mutants affected in siRNAs biogenesis, questioning the role of siRNAs in the control of germinal and, therefore, transgenerational transposon mobility[11].

We showed previously that a temperature shift applied to 1-week-old seedlings transiently destabilized transcriptional gene silencing (TGS) at loci residing within constitutive heterochromatin where TGS was re-established during the next 24 h[14]. A notable exception was a *Ty1/copia*-type long terminal repeat (LTR) retrotransposon family (*ATCOPIA78*), which retained high levels of transcripts two days later[14]. This was also observed in older plants (21 days) subjected to raised temperatures[15]. In the genome of the Columbia accession, *ATCOPIA78* consists of eight members (Supplementary Fig. 1a), hereafter referred to as *ONSEN*, of which three have identical LTR sequences, indicating recent transposition (Supplementary Fig. 1b).

By northern blotting, we compared *ONSEN* transcripts in seedlings subjected to a temperature shift of 24 h at 6 °C followed by 24 h at 37 °C (hereafter called heat stress (HS)) to transcripts of seedlings subjected to a control stress (CS) of 24 h at 6 °C followed by 24 h at 21 °C (Fig. 1a). *ONSEN* transcripts were detected in HS plants directly after the stress treatment and for up to 3 days of recovery at 21 °C (HS+3). The longest RNA found corresponded to the full-length transposon (Fig. 1a), whereas smaller RNAs appeared to belong to aberrant RNAs often associated with transcriptionally activated retroelements[9,16]. Full-length *ONSEN* transcripts were not observed in plants subjected to CS or in non-stressed plants (Fig. 1a and data not shown).

To examine further the specificity of TGS release and to determine possible epigenetic mechanisms involved in *ONSEN* control, we tested RNAs of plants treated with DNA methylation inhibitor 5-azacytidine (AzaC, Fig. 1b) and of *ddm1* mutant plants (Fig. 1c). Neither AzaC treatment nor *ddm1* mutation was effective for transcriptional activation of *ONSEN*, indicating that a reduction of DNA methylation is not sufficient for releasing *ONSEN* silencing. Furthermore, we applied HS and CS treatments to mutants compromised in epigenetic regulation (Fig. 1c, d). We examined *ddm1* mutants (Fig. 1c) and mutants affected in siRNA biogenesis (Fig. 1d): *nrpd1* (ref. 3), impaired in plant-specific RNA polymerase IV (PolIV); *nrpd2* (ref. 17), impaired in the common subunit of RNA PolIV and PolV; *rdr2* (ref. 18), impaired in RNA-dependent RNA polymerase 2; and *dcl3* (ref. 18), mutated in Dicer-like 3. We also challenged the *suvh2* (ref. 19) mutant (Fig. 1d), which is deficient in a putative histone 3 lysine 9 methyltransferase. *ONSEN* transcripts were only observed in RNA samples after HS but not after CS treatment (Fig. 1a, c, d). Their levels were not affected by the *ddm1* mutation (Fig. 1c). In contrast, HS-induced accumulation of *ONSEN* RNA was significantly higher in *nrpd1, nrpd2, rdr2, dcl3* and *suvh2* mutants (Fig. 1d). During the recovery period following stress, *ONSEN* transcripts diminished and after 10 days (HS+10) the full-length RNAs of the transposon were not detected on northern blots of all genotypes tested (Fig. 1a and Supplementary Fig. 2). These results indicated that siRNA-mediated regulation is responsible for the restriction of *ONSEN* transcript levels after HS, but is not involved in resilencing during the recovery period.

As PolIV is crucial for the biogenesis of the majority of 24-nucleotide siRNAs[4], we compared the levels of *ONSEN*-specific siRNAs in wild-type and *nrpd1* plants in relation to the HS-induced accumulation of its transcripts. Noticeably, directly after HS, when *ONSEN* transcript levels were highest, siRNA levels were low and increased only after 1 day of recovery (Fig. 1a, e, f). These siRNAs appeared in both wild-type and *nrpd1* plants and were mainly of the 21-nucleotide siRNA class (Fig. 1e, f), which is thought to direct cleavage of corresponding messenger RNAs. Although levels of 21-nucleotide siRNAs were significantly higher in *nrpd1* mutant plants than in the wild type, massive

[1] Department of Plant Biology, University of Geneva, Sciences III, 30 Quai Ernest-Ansermet, CH-1211 Geneva 4, Switzerland. †Present addresses: Division of Biological Sciences, Graduate School of Science, Hokkaido University, 060-0810 Sapporo, Japan (H.I.); Botanical Institute, Hebelstrasse 1, CH-4056 Basel, Switzerland (E.B.); Institut de Recherche pour le Développement (IRD), UMR DIADE, Plant Diversity, Adaptation and Development Laboratory, Université Montpellier 2, 911 Avenue Agropolis, 34394 Montpellier, France (M.M.); Centre National de la Recherche Scientifique (CNRS), UMR 6247 - GReD - INSERM U 931, Clermont Université, 24 avenue des Landais, BP 80026, 63171 Aubière, France (I.V.).
*These authors contributed equally to this work.

©2011 Macmillan Publishers Limited. All rights reserved

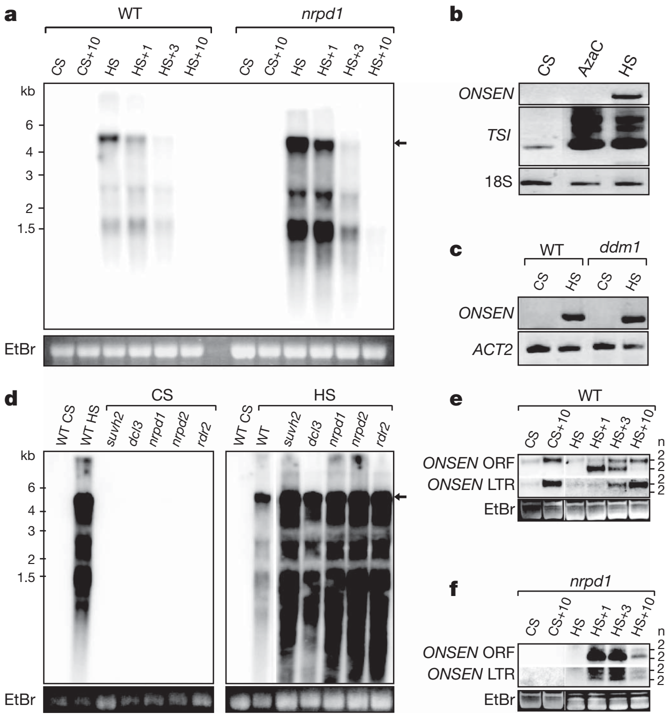


Figure 1 | Heat-stress induced *ONSEN* transcription. **a**, Northern blot revealing *ONSEN* transcripts in wild-type (WT) and *nrpd1* seedlings subjected to HS and after recovering from HS for 1, 3 and 10 days (HS+1, HS+3, HS+10); CS plants were subjected to the control stress (CS). The arrow marks the full-length *ONSEN* transcript. An ethidium bromide (EtBr)-stained gel is shown as loading control. **b**, Detection of *ONSEN* transcripts after AzaC and HS treatments by semiquantitative reverse transcription followed by PCR (RT–PCR) in wild-type plants. *TRANSCRIPTIONALLY SILENT INFORMATION* (*TSI*) was used as a positive control for the activation of heterochromatic transcription[14]; 18S ribosomal RNA was used as an internal control. **c**, Levels of *ONSEN* transcripts in *ddm1* mutant plants subjected to CS or HS and quantified by RT–PCR with *ACTIN2* transcripts (*ACT2*) as an internal control. **d**, Northern blots showing *ONSEN* transcript levels in selected mutants (marked above each lane) subjected to CS or HS. The CS blot was overexposed for possible detection of low levels of *ONSEN* transcripts. An EtBr-stained gel below is shown as a loading control. **e**, **f**, Northern blots of *ONSEN* siRNAs derived from open-reading frame (ORF) or LTR regions accumulating in wild-type (**e**) and *nrpd1* mutant plants (**f**). An EtBr-stained gel is shown as a loading control. nt, nucleotide. See Methods for probe information.

amounts of *ONSEN* transcripts were observed. Thus, this siRNA class was not able to prevent the accumulation of transposon-derived mRNA. However, the high background smear visible on northern blots (Fig. 1a, d) and even more apparent on overexposed blots of RNA isolated from mutants impaired in the siRNA pathway 10 days after HS (Supplementary Fig. 2) may be indicative of *ONSEN* transcript degradation. In the course of HS recovery (at 3 and 10 days), 24-nucleotide siRNAs highly accumulated in wild-type but not in *nrpd1* plants, and were especially abundant for the LTR regions of *ONSEN* (Fig. 1e, f). However, the contribution of 24-nucleotide siRNAs to resilencing at *ONSEN* loci during recovery is not clear, as they accumulate also in CS plants grown for an additional 10 days after CS treatment (Fig. 1e). Therefore their levels seem not to be related to the HS treatment. Moreover, *ONSEN* resilencing also occurred in *nrpd1,* where they were mostly absent (Fig. 1f).

Detection of *ONSEN* full-length transcripts, potentially able to serve as templates for reverse transcription, prompted us to examine the DNA of *Arabidopsis* subjected to HS. By Southern blot analysis we detected a significant increase in *ONSEN* copy number and observed a banding pattern indicative of the presence of two forms of extrachromosomal transposon copies, one linear reflecting the 2.8-kb fragment and one circular containing a single LTR consistent with the 4.5-kb fragment (Fig. 2a, left). Linear extrachromosomal forms are capable of chromosomal integration, in contrast to the circular forms that have been considered as by-products of retroelement replication[20,21]. In *nrpd1* and other mutant plants affected in the siRNA pathway, the abundance of HS-induced *ONSEN* DNA was significantly higher than in wild type (Fig. 2a, left). Noticeably, similarly high levels were observed in HS-treated *suvh2* mutant plants (Fig. 2a, left), which are known to exhibit wild-type levels of siRNAs[22]. After 10 days of recovery, *ONSEN* extrachromosomal DNA was still at a relatively high level but almost exclusively in the linear form (Fig. 2a, right).

Real-time quantitative polymerase chain reaction (qPCR) during HS and after subsequent recovery was performed to determine the kinetics of *ONSEN* DNA accumulation in wild-type plants and in a representative siRNA-biogenesis mutant (*nrpd1*) and to examine possible changes in copy number due to chromosomal integration events (Fig. 2b). Over the first 4 h of the temperature shift from 6 °C to 37 °C, the abundance of *ONSEN* DNA did not change. However, after 6 h, *ONSEN* copy number increased significantly from 8 endogenous copies to more than 30 in *nrpd1* but not in the wild type (Fig. 2b). After 12 h of HS, *ONSEN* copy number had increased in the wild type to more than 25 and in the *nrpd1* mutant to more than 160. The maximal copy numbers of over 50 for the wild type and over 500 for *nrpd1* mutants were reached 12 h and 24 h after HS, respectively. The HS-induced increase in *ONSEN* DNA seemed to be biphasic and this was especially pronounced in *nrpd1* plants (Fig. 2b). So far, we have no explanation for this biphasic accumulation but it may be related to stress-triggered synchronization of the retroelement replication cycle.

During 20–30 days of subsequent growth of both wild-type and *nrpd1* plants *ONSEN* copy number gradually decreased, reaching the initial number of the Columbia accession (Fig. 2b), consistent with the absence of or only sporadic chromosomal integration events. To examine whether new somatic integrations occurred, we performed transposon display on plants grown for 40 days after HS (Supplementary Fig. 3a). New *ONSEN* insertions were not detected in the genomic DNA of either wild-type or *nrpd1* mutant plants, consistent with the qPCR results. However, we can not exclude the possibility of rare transposition events occurring late in vegetative development leading to only small sectors with new insertions.

It has been suggested that transgenerational transposon mobility is suppressed during gametophyte formation by siRNAs[10,23]. However, there is no evidence based on germinal transposition events to support this hypothesis. By transposon display and Southern blot hybridization,

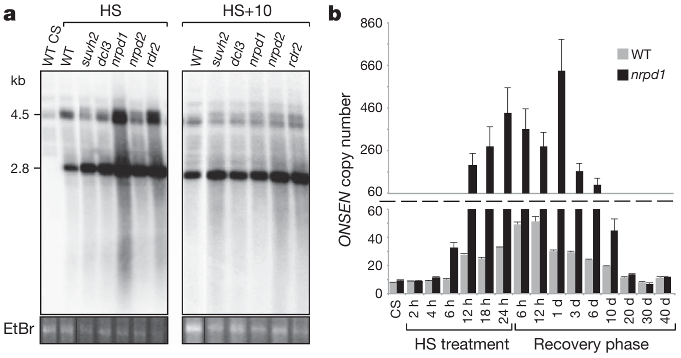


Figure 2 | Accumulation of *ONSEN* extrachromosomal DNA. **a**, Southern blot of PsiI-digested DNA isolated from HS-treated seedlings of wild type and selected mutants, directly after HS (left) or after 10 days of recovery (HS+10, right) and hybridized with an *ONSEN*-specific probe (see Methods). A 2.8-kb band is expected for the extrachromosomal linear form of *ONSEN* (Supplementary Fig. 1b). **b**, The kinetics of *ONSEN* DNA accumulation for wild type (grey) and *nrpd1* (black) measured by qPCR during and after HS treatment (mean ± s.e.m., $n = 3$ biological repetitions).

©2011 Macmillan Publishers Limited. All rights reserved

we examined genomic DNA from the progeny of self-fertilized wild-type and *nrpd1* plants subjected to HS and CS for new *ONSEN* insertions. Transposon movement was not detected in the offspring of either *nrpd1* or wild-type plants subjected to CS, or in HS-treated wild type (Fig. 3a and Supplementary Fig. 3b). However, a surprisingly high frequency of retrotransposition was recorded in the progeny of *nrpd1* mutant plants subjected to HS at the seedling stage (Fig. 3a and Supplementary Fig. 3b, c). Furthermore, the patterns of new *ONSEN* insertions in sibling seedlings derived from a single plant were found to differ in each individual examined, indicating that transposition occurred either before gametogenesis, during gametogenesis, after fertilization, or any combination therein (Fig. 3a and Supplementary Fig. 3b). To distinguish between these alternatives, we analysed *nrpd1* progeny plants derived from seeds of different flowers of the same progenitor (Fig. 3b). We found that patterns of new insertions differed entirely between progeny derived from different flowers. However, within the same flower we found common transposition patterns indicating somatic movement of *ONSEN* during flower development. Moreover, we were not able to find any new and unique *ONSEN* insertions specific to a single plant (Fig. 3b). Therefore all transposition events revealed in the sixteen progeny plants derived from two different flowers must have occurred before the differentiation of male and female gametophytes. Therefore, siRNA-mediated control of retrotransposon movement is not restricted to the gametophytic phase as it has been postulated[10–13].

To define better the roles of sporophytic and gametophytic 24-nucleotide siRNAs in suppressing stress-induced transgenerational retrotransposition, we examined heterozygote *nrpd1* plants subjected to HS treatment. In these plants the biogenesis of siRNAs is unaffected in the sporophyte, but is deficient in 50% of the male and female gametophytes. We compared the progeny of homozygous *nrpd1* mutant plants subjected to HS with homozygous *nrpd1* mutant

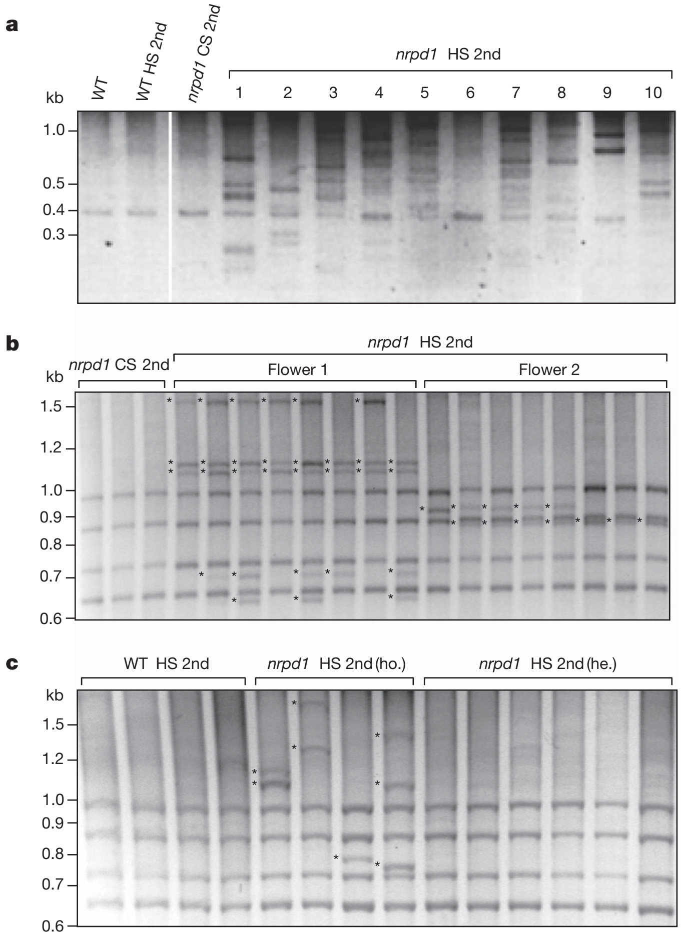


Figure 3 | Burst of *ONSEN* transposition in the progeny of HS-treated *nrpd1* plants. **a**, Transposon display (using primer Copia78 3′ LTR, Supplementary Table 2) detecting new *ONSEN* insertions. Numbers above the lanes of *nrpd1* HS 2nd (second generation) represent 10 individual plants that are siblings derived from bulk-harvested seeds of one *nrpd1* plant that was HS-treated as a 7-day-old seedling. **b**, Transposon display (using primer ONS_312_R, Supplementary Table 2) detecting new *ONSEN* insertions. Sixteen *nrpd1* HS 2nd plants are derived from two flowers of a single HS-treated *nrpd1* plant (flower 1 and flower 2 represented by eight plants each). Asterisks mark new *ONSEN* insertions. **c**, Transposon display (with primer as in **b**) revealing new *ONSEN* insertions in the progeny of HS-treated *nrpd1* homozygote mutant plants (*nrpd1* HS 2nd (ho.)) but not in the progeny of wild-type HS-treated plants (WT HS 2nd) or in *nrpd1* homozygote mutant progeny of HS-treated *nrpd1* heterozygote mutants (*nrpd1* HS 2nd (he.)).

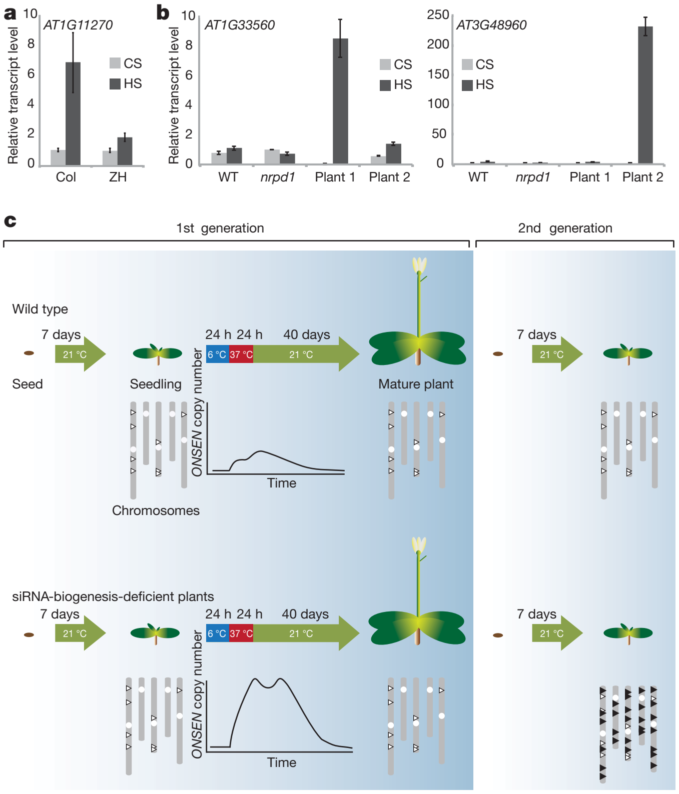


Figure 4 | Acquisition of heat-responsive transcription by *ONSEN*-containing loci and model of *ONSEN* activation. **a**, Effect of *ONSEN* natural insertion polymorphism on gene regulation. *AT1G11270* transcript levels after CS (grey) and HS (black) treatments were analysed by qRT–PCR in Columbia (Col) and Zürich (ZH) accessions in which *ONSEN* is present or absent at this locus, respectively (mean ± s.e.m., $n = 3$, values relative to *ACT2*). **b**, HS-mediated transcriptional regulation of two different genes acquiring new *ONSEN* insertion in the course of our experiments. Two progeny plants of HS-treated *nrpd1* homozygote mutant plant, named plant 1 and plant 2, were selected for displaying new homozygous *ONSEN* insertion at two distinct loci, *AT1G33560* and *AT3G48960*, respectively. Acquired transcriptional responses to HS of the affected genes were revealed by qRT–PCR (legend and values as in **a**). **c**, Summary of experimental results illustrating the role of the siRNA pathway in transgenerational control of *ONSEN* mobility. Upper part of the figure represents wild-type control of *ONSEN* activity and lower part illustrates uncontrolled accumulation of *ONSEN* copy number in siRNA-biogenesis-deficient plants. The graphs under the arrows illustrate the kinetics of *ONSEN* DNA accumulation on HS treatment. The open triangles on five *Arabidopsis* chromosomes represent eight endogenous *ONSEN* copies in the Columbia accession. The black triangles illustrate new *ONSEN* insertions found in the second generation. White circles on the chromosomes specify the location of the centromeres.

©2011 Macmillan Publishers Limited. All rights reserved

segregants derived from HS-treated *nrpd1* heterozygotes. A high frequency of retrotransposition was only observed in progeny of homozygous *nrpd1* mutant plants (Fig. 3c). These results are consistent with the involvement of 24-nucleotide siRNAs in either erasing 'stress memory' during somatic growth and/or suppressing retrotransposition in flower tissues, rather than with epigenetic control of retrotransposon movement during gametogenesis.

To define better the molecular mechanism controlling *ONSEN* transposition primed by HS, we analysed progenies of further HS-treated plants compromised in epigenetic regulation. Because DDM1 and KRYPTONITE (KYP, histone H3 lysine 9 methyltransferase) were previously implicated in transposition control of a related family of retrotransposons[6,9], we subjected both mutants to HS and examined the progeny by transposon display. Retrotransposition was not observed (Supplementary Fig. 4). Despite detecting high levels of transposon transcripts in *suvh2* mutants after HS (Fig. 1d) and a significant increase in *ONSEN* copy number (Fig. 2a), no retrotransposition events were found in the next generation (Supplementary Fig. 4). We next investigated mutants deficient in siRNA biogenesis (*nrpd2, rdr2* and *dcl3*). Transposition events were observed in *nrpd2* and *rdr2* (Supplementary Fig. 4), further indicating that biogenesis of siRNAs is crucial for preventing transgenerational mobility of *ONSEN*. As *dcl3* is essential for 24-nucleotide siRNA biogenesis, we predicted that there would be new *ONSEN* insertions in the progeny of HS-treated *dcl3* plants; in fact, no new insertions were detected (Supplementary Fig. 4). Therefore, although DCL3 clearly restricts the levels of *ONSEN* transcripts after HS, it is dispensable for the control of transgenerational transposition. This hints at two steps in *ONSEN* control: restraining levels of its transient transcription/reverse transcription and suppression of transgenerational transposition. As only the first requires DCL3 and SUVH2, the two control steps seem to be, at least in part, mechanistically independent. Given the functional redundancy of dicer-like (DCL) proteins in *Arabidopsis*[24], DCL3 may be substituted possibly by another DCL protein(s) at the second control step. It is also possible that transgenerational control of retrotransposition can occur without the involvement of dicer-like activities, as has been described in animals[25].

To determine whether *ONSEN* has preferential insertion targets, we characterized 11 new insertion sites and concluded that, although the retroelement inserted genome wide (Supplementary Fig. 5), it showed a clear preference for transcribed gene regions (all 11 insertions), with a further preference for exons (10 insertions) (Supplementary Table 1). Moreover, 2 of 11 insertions were homozygous (data not shown), which is consistent with retrotransposon movement during flower development but before the differentiation of anthers and carpels.

It has been postulated that a burst of transposition helped to shape plant genomes[26,27] and to modify their transcriptional responses[28]. Interestingly, a gene in the Columbia accession harbouring a natural insertion of *ONSEN* was identified as being heat responsive[29]. To determine the physiological relevance of this observation we analysed heat responsiveness of this gene in the Zürich accession where *ONSEN* is absent at this location (Fig. 4a). Indeed, HS-induced transcriptional activation in the Columbia accession was much more pronounced than in the Zürich accession (Fig. 4a). We determined also whether our experimentally induced retrotransposition events, in the second generation of *nrpd1* HS-treated plants, had an impact on the transcriptional regulation of endogenous loci harbouring new *ONSEN* insertions. We examined the heat-stress response of two such genes and showed that they became heat responsive when compared to wild-type or *nrpd1* first-generation plants (Fig. 4b). Therefore, it can be predicted that after our experimental burst of *ONSEN* transposition different subsets of genes in various progeny plants will acquire such regulatory properties. Now, having established an environmentally inducible system of transgenerational retrotransposition and having revealed the molecular and developmental mechanisms of its control (Fig. 4c), we are in a position to reproduce retrotransposon bursts in a controlled fashion and to determine their adaptive and/or damaging power.

METHODS SUMMARY

Plant material. All mutants used in this study (*dcl3-1* (ref. 18), *ddm1-2* (ref. 30), *nrpd1a-3* (ref. 3), *nrpd2a-2/2b-1* (ref. 17), *rdr2-2* SALK_059661 (ref. 18), *suvh2* SALK_079574 (ref. 19)) are in the Columbia (Col-0) background.

Stress treatment. Plants were grown in ½ MS medium (0.8% agar, 1% sucrose) in a Percival CU-22L chamber at 21 °C with 12 h light (140 μmol m^{-2} s^{-1}) and 12 h dark. After CS or HS treatment (see text), plants were grown at 21 °C in long-day conditions (16 h light). To analyse the progeny of CS- or HS-treated seedlings, plants were transplanted 10 days after CS/HS to soil and grown under long-day conditions.

RNA, DNA, and transposon display analyses. See Methods for details.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 July 2010; accepted 13 January 2011.
Published online 13 March 2011.

1. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nature Rev. Genet.* **8,** 272–285 (2007).
2. Lisch, D. Epigenetic regulation of transposable elements in plants. *Annu. Rev. Plant Biol.* **60,** 43–66 (2008).
3. Herr, A. J., Jensen, M. B., Dalmay, T. & Baulcombe, D. C. RNA polymerase IV directs silencing of endogenous DNA. *Science* **308,** 118–120 (2005).
4. Mosher, R. A., Schwach, F., Studholme, D. & Baulcombe, D. C. PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. *Proc. Natl Acad. Sci. USA* **105,** 3145–3150 (2008).
5. Miura, A. *et al.* Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis. Nature* **411,** 212–214 (2001).
6. Tsukahara, S. *et al.* Bursts of retrotransposition reproduced in *Arabidopsis. Nature* **461,** 423–426 (2009).
7. Johannes, F. *et al.* Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet.* **5,** e1000530 (2009).
8. Reinders, J. *et al.* Compromised stability of DNA methylation and transposon immobilization in mosaic *Arabidopsis* epigenomes. *Genes Dev.* **23,** 939–950 (2009).
9. Mirouze, M. *et al.* Selective epigenetic control of retrotransposition in *Arabidopsis. Nature* **461,** 427–430 (2009).
10. Slotkin, R. K. *et al.* Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* **136,** 461–472 (2009).
11. Mosher, R. A. *et al.* Uniparental expression of PolIV-dependent siRNAs in developing endosperm of *Arabidopsis. Nature* **460,** 283–286 (2009).
12. Hsieh, T.-F. *et al.* Genome-wide demethylation of *Arabidopsis* endosperm. *Science* **324,** 1451–1454 (2009).
13. Gehring, M., Bubb, K. L. & Henikoff, S. Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science* **324,** 1447–1451 (2009).
14. Tittel-Elmer, M. *et al.* Stress-induced activation of heterochromatic transcription. *PLoS Genet.* **6,** e1001175 (2010).
15. Pecinka, A. *et al.* Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in *Arabidopsis. Plant Cell* **22,** 3118–3129 (2010).
16. Hirochika, H., Okamoto, H. & Kakutani, T. Silencing of retrotransposons in *Arabidopsis* and reactivation by the ddm1 mutation. *Plant Cell* **12,** 357–369 (2000).
17. Onodera, Y. *et al.* Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120,** 613–622 (2005).
18. Xie, Z. *et al.* Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol.* **2,** e104 (2004).
19. Naumann, K. *et al.* Pivotal role of AtSUVH2 in heterochromatic histone methylation and gene silencing in *Arabidopsis. EMBO J.* **24,** 1418–1429 (2005).
20. Feuerbach, F., Drouaud, J. & Lucas, H. Retrovirus-like end processing of the tobacco *Tnt1* retrotransposon linear intermediates of replication. *J. Virol.* **71,** 4005–4015 (1997).
21. Hirochika, H. & Otsuki, H. Extrachromosomal circular forms of the tobacco retrotransposon *Tto1. Gene* **165,** 229–232 (1995).
22. Johnson, L. M., Law, J. A., Khattar, A., Henderson, I. R. & Jacobsen, S. E. SRA-domain proteins required for DRM2-mediated *de novo* DNA methylation. *PLoS Genet.* **4,** e1000280 (2008).
23. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Rev. Genet.* **11,** 204–220 (2010).
24. Vaucheret, H. Plant ARGONAUTES. *Trends Plant Sci.* **13,** 350–358 (2008).
25. Vagin, V. V. *et al.* A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313,** 320–324 (2006).
26. Piegu, B. *et al.* Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **16,** 1262–1269 (2006).
27. Ammiraju, J. S. S. *et al.* Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza. Plant J.* **52,** 342–351 (2007).
28. Naito, K. *et al.* Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* **461,** 1130–1134 (2009).
29. Lim, C. J. *et al.* Gene expression profiles during heat acclimation in *Arabidopsis thaliana* suspension-culture cells. *J. Plant Res.* **119,** 373–383 (2006).
30. Vongs, A., Kakutani, T., Martienssen, R. A. & Richards, E. J. *Arabidopsis thaliana* DNA methylation mutants. *Science* **260,** 1926–1928 (1993).

©2011 Macmillan Publishers Limited. All rights reserved

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank all members of the J.P. laboratory and R. L. Fisher for discussions, L. Broger, M. Freyre, J. Nicolet, C. Mégies for technical assistance and P. King and J. Reinders for editing of the manuscript. This work was supported by grants from the Swiss National Science Foundation (31003A-125005) and the European Commission through the AENEAS collaborative project (FP7 226477), the RECBREED consortium (FP7 227190) and The Epigenome Network of Excellence (FP6 503433).

Author Contributions H.I., E.B., H.G., M.M. and J.P. conceived the study. H.I., E.B., H.G., M.M. and I.V. performed the experiments. J.P. wrote the paper with contributions from E.B. and M.M.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.P. (jerzy.paszkowski@unige.ch).

©2011 Macmillan Publishers Limited. All rights reserved

METHODS

Plant material. All mutants used in this study (*dcl3-1* (ref. 18), *ddm1-2* (ref. 30), *nrpd1a-3* (ref. 3), *nrpd2a-2/2b-1* (ref. 17), *rdr2-2* SALK_059661 (ref. 18), *suvh2* SALK_079574 (ref. 19)) are in the Columbia (Col-0) background.

Stress treatment. Plants were grown in ½ MS medium (0.8% agar, 1% sucrose) in a Percival CU-22L chamber at 21 °C with 12 h light (140 μmol $m^{-2} s^{-1}$) and 12 h dark. After CS or HS treatment (see text), plants were grown at 21 °C in long-day conditions (16 h light). To analyse the progeny of CS- or HS-treated seedlings, plants were transplanted 10 days after CS/HS to soil and grown under long-day conditions.

RNA analysis. RNA was isolated from aerial parts of around 20 plants and northern blots, RT–PCR and siRNA analyses were carried out as described previously[9]. Full-length transcripts were detected with *ONSEN*-specific probe A, and siRNAs were detected with *ONSEN*-specific probe B (LTR region) or probe C (ORF). qRT–PCR analyses were performed using the Quantifast Multiplex PCR Kit (Qiagen). RNA levels were determined using TaqMan assays (qPCR thermocycler 7900HT, Applied Biosystems) and normalized using *ACTIN2*. PCR conditions were 95 °C for 5 min followed by 45 cycles alternating 45 s at 95 °C and 45 s at 60 °C. Probe localization and primer details are given in Supplementary Fig. 1b and Supplementary Table 2, respectively.

DNA analysis. Aerial parts from around 20 plants were collected and DNA was isolated with MiniPrep Kit (Qiagen) following the manufacturer's recommendations. Southern blots were performed as described previously[9] using *ONSEN*-specific probe (probe C, see Supplementary Fig. 1b and Supplementary Table 2). For qPCR analysis of *ONSEN* DNA copies, the Quantifast Multiplex PCR Kit (Qiagen) was used and *ACTIN2* was used to normalize DNA levels. DNA copy number was determined using TaqMan assays performed in the qPCR thermocycler 7900HT (Applied Biosystems) in a final volume of 10 μl. PCR conditions were 95 °C for 5 min followed by 45 cycles alternating 45 s at 95 °C and 45 s at 60 °C (primer details in Supplementary Table 2).

Transposon display. A simplified transposon display method based on the GenomeWalker Universal kit (ClontechLaboratories) was developed for library construction, with the following modifications. Genomic DNA (300 ng) was digested overnight with the blunt cutting DraI restriction enzyme (Promega) in a final volume of 50 μl, using a tenfold enzyme excess compared with the manufacturer's recommendations. After digestion, DNA fragments were purified on a PCR purification column (Qiagen) following the manufacturer's instructions and eluted into 20 μl; 5 μl was used for overnight ligation at 16 °C in 16 μl with GenomeWalker adaptors. After ligation, DNA was diluted 20-fold and 1 μl used as a template for the PCR reaction. PCR were performed using a primer specific for *ONSEN* (Copia78 3′ LTR or ONS_312_R) and a primer specific for the adaptor (GenWalk_AP1). PCR conditions were 5 min at 95 °C, followed by 33 cycles of 30 s at 94 °C, 30 s at 58 °C, 1 min at 72 °C; and a final elongation step of 7 min at 72 °C. PCR products were separated on 3% agarose gels. Primer details are given in Supplementary Table 2.

©2011 Macmillan Publishers Limited. All rights reserved

doi:10.1038/nature09819

A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression

Kevin C. Wang[1,2], Yul W. Yang[1]*, Bo Liu[3]*, Amartya Sanyal[4], Ryan Corces-Zimmerman[1], Yong Chen[5], Bryan R. Lajoie[4], Angeline Protacio[1], Ryan A. Flynn[1], Rajnish A. Gupta[1], Joanna Wysocka[6], Ming Lei[5], Job Dekker[4], Jill A. Helms[3] & Howard Y. Chang[1]

The genome is extensively transcribed into long intergenic noncoding RNAs (lincRNAs), many of which are implicated in gene silencing[1,2]. Potential roles of lincRNAs in gene activation are much less understood[3–5]. Development and homeostasis require coordinate regulation of neighbouring genes through a process termed locus control[6]. Some locus control elements and enhancers transcribe lincRNAs[7–10], hinting at possible roles in long-range control. In vertebrates, 39 *Hox* genes, encoding homeodomain transcription factors critical for positional identity, are clustered in four chromosomal loci; the *Hox* genes are expressed in nested anterior-posterior and proximal-distal patterns colinear with their genomic position from 3′ to 5′of the cluster[11]. Here we identify *HOTTIP*, a lincRNA transcribed from the 5′ tip of the *HOXA* locus that coordinates the activation of several 5′ *HOXA* genes *in vivo*. Chromosomal looping brings *HOTTIP* into close proximity to its target genes. *HOTTIP* RNA binds the adaptor protein WDR5 directly and targets WDR5/MLL complexes across *HOXA*, driving histone H3 lysine 4 trimethylation and gene transcription. Induced proximity is necessary and sufficient for *HOTTIP* RNA activation of its target genes. Thus, by serving as key intermediates that transmit information from higher order chromosomal looping into chromatin modifications, lincRNAs may organize chromatin domains to coordinate long-range gene activation.

We examined chromosome structure and histone modifications in human primary fibroblasts derived from several anatomic sites[12], and found distinctive differences in the *HOXA* locus. High throughput chromosome conformation capture (5C)[13] across *HOXA* revealed that its higher order structure is dependent on positional identity. In anatomically distal cells (for example, foreskin and foot fibroblasts), we detected abundant chromatin interactions within the transcriptionally active 5′ *HOXA* locus (with reference to the directions of transcription of constituent *Hox* genes), pointing to a compact and looped conformation. In contrast, no long-range chromatin interactions are detected within the transcriptionally silent 3′ *HOXA* which seems largely linear (Fig. 1a). Strikingly, anatomically proximal cells (for example, lung fibroblasts) have the diametrically opposite pattern. The ON and OFF states of *Hox* and other key developmental genes are maintained by the MLL/Trithorax (Trx) and polycomb group (PcG) proteins, which mediate trimethylation of histone H3 lysine 4 (H3K4me3) to activate genes or lysine 27 (H3K27me3) to repress genes[14]. The portions of *HOXA* in tight physical interaction are marked by broad domains of H3K4me3, whereas H3K27me3 marks the physically extended and transcriptional silent regions (Fig. 1a).

On the very 5′ and 3′ edges of the two respective interaction clusters are two lincRNA loci that exhibit distinct chromatin modifications. The 3′element has been previously identified as the myelopoiesis-associated lincRNA *HOTAIRM1* (ref. 15). The 5′ element, for which we suggest the name *HOTTIP* for '*HOXA* transcript at the distal tip', exhibits bivalent H3K4me3 and H3K27me3, a histone modification pattern associated with poised regulatory sequences[16]. Comparison with RNA polymerase II occupancy and RNA expression showed that the bivalent H3K4me3 and H3K27me3 modifications on *HOTTIP* gene do not require *HOTTIP* transcription, but transcription of *HOTTIP* is associated with increased H3K4me3 and decreased H3K27me3 (Fig. 1a, left). Chromatin immunoprecipitation (ChIP) analysis confirmed that the *HOTTIP* gene is occupied by both polycomb repressive complex 2 (PRC2) and MLL complex, consistent with the bivalent histone marks (Supplementary Fig. 1a).

HOTTIP transcription yields a 3,764-nucleotide, spliced and polyadenylated lincRNA that initiates ~330 bases upstream of *HOXA13*. Only the strand antisense to *HOXA* genes is transcribed (Supplementary Fig. 1b). Genes near the 5′ end of each *HOX* cluster tend to be expressed in more posterior and/or distal anatomical locations. Consistent with its genomic location 5′ to *HOXA13*, *HOTTIP* is expressed in distal and/or posterior anatomic sites (Fig. 1b). *In situ* hybridization of developing mouse and chick embryos confirmed that *HOTTIP* is expressed in posterior and distal sites *in vivo*, indicating a conserved expression pattern from development to adulthood (Fig. 1c and Supplementary Fig. 1c). Even in distal cells where *HOTTIP* is expressed, its RNA level is very low and estimated to be ~0.3 copies per cell (Supplementary Fig. 2).

We employed small interfering RNAs (siRNAs) to knock down *HOTTIP* RNA in fibroblasts from a distal anatomic site (foreskin), and examined expression of 5′ *HOXA* genes by quantitative reverse transcription PCR. Notably, *HOTTIP* RNA knockdown abrogated expression of distal *HOXA* genes across 40 kilobases with a trend dependent on the distance to *HOTTIP. T*he strongest blockade was observed for *HOXA13* and *HOXA11*, with progressively less severe effects on *HOXA10*, *HOXA9* and *HOXA7* (Fig. 2a). The effect on gene transcription appeared to be unidirectional, as there were no appreciable changes in the levels of *EVX1*, located ~40 kilobases 5′ of the *HOXA* cluster (data not shown). *HOTTIP* knockdown did not affect expression of the highly homologous *HOXD* genes, other control genes, nor induce antisense transcription at its own locus (Fig. 2b, Supplementary Fig. 3a). Several independent siRNAs targeting *HOTTIP* yielded similar results (Supplementary Fig. 3b). These results indicate that *HOTTIP* RNA is necessary to coordinate activation of 5′ *HOXA* genes.

We next addressed the function of *HOTTIP* RNA *in vivo* in the developing chick limb bud (Fig. 2c). Whereas prior genetic studies of noncoding RNAs (ncRNAs) involved deletion or insertion into the gene locus[17], we wished to distinguish the functions of *HOTTIP* RNA from its corresponding DNA element. *HOTTIP* gene can nucleate H3K4 and H3K27 methylation independent of transcription (Fig. 1a),

[1]Howard Hughes Medical Institute, Program in Epithelial Biology, Stanford University School of Medicine, Stanford, California 94305, USA. [2]Department of Dermatology, University of California San Francisco (UCSF), San Francisco, California 94115, USA. [3]Department of Surgery, Stanford University School of Medicine, Stanford, California 94305, USA. [4]Program in Gene Function and Expression, Dept. of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA. [5]Howard Hughes Medical Institute, Department of Biological Chemistry, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. [6]Department of Chemical and Systems Biology, Stanford University School of Medicine, Stanford, California 94305.

*These authors contributed equally to this work.

©2011 Macmillan Publishers Limited. All rights reserved

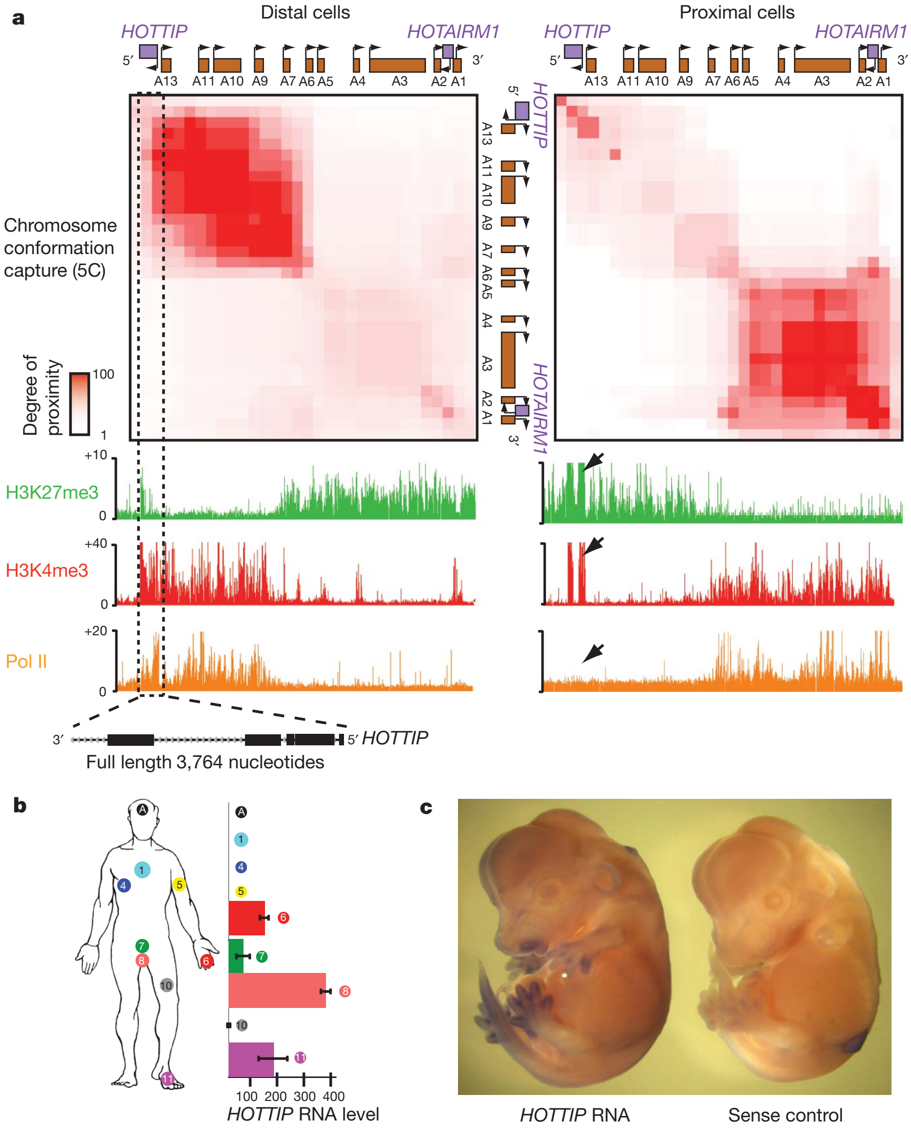


Figure 1 | *HOTTIP* is a lincRNA transcribed in distal anatomic sites. **a**, Chromatin state map of distal versus proximal cells. Top panels, chromosome conformation capture-carbon copy (5C) analysis of distal (foreskin) and proximal (lung) human fibroblasts. Heat map representations (generated by my5C, ref. 30) of 5C data (bin size 30 kb, step size 3 kb) for *HOXA* in foreskin and lung fibroblasts. Red intensity of each pixel indicates relative interaction between the two points on the genomic coordinates. The diagonal represents frequent *cis* interactions between regions located in close proximity along the linear genome. 5C signals that are away from the diagonal represent long-range looping interactions. Bottom panels, chromatin occupancy across *HOXA*. *x*-axis is genomic coordinate; *y*-axis depicts occupancy of the indicated histone marks or protein (ChIP/input). Box and arrows highlight chromatin states of *HOTTIP* gene. **b**, *HOTTIP* RNA expression in primary human fibroblasts from 11 anatomic sites. Means ± s.d. are shown ($n = 2$). **c**, *In situ* hybridization of *HOTTIP* RNA in E13.5 mouse embryo.

and the precise genomic distance between upstream enhancer elements and *Hox* genes is critical for their proper colinear activation[17]. Therefore, we used RNA interference (RNAi) in chick embryos, where replication-competent retroviruses can deliver short-hairpin RNAs (shRNAs) with high penetrance and precise spatiotemporal control[18] (Supplementary Fig. 4). In the limb bud, 5′ *HoxA* genes are transcribed in a nested pattern along the proximal-distal axis[19]. In this tissue, *HoxA* function is highly redundant with that of the *HoxD* locus, which allowed us to assess altered *HoxA* expression patterns without major changes in anatomic landmarks[20]. We injected retroviruses carrying shRNAs against chick *HOTTIP* into upper limb buds of stage 13 chicks; RT–PCR and *in situ* hybridization were performed on both control and knockdown samples after 2–4 days. Knockdown of *HOTTIP* RNA by two independent shRNAs in limb buds decreased expression of *HoxA13*, *HoxA11* and *HoxA10*—again with a graded impact depending on genomic proximity to *HOTTIP* gene. Vector control or an shRNA that fails to deplete *HOTTIP* RNA had little effect on *Hox* gene expression (Fig. 2d). *In situ* hybridization on whole embryos (Fig. 2e) and sections (Supplementary Fig. 5) revealed that *HOTTIP* RNA most strongly affects *HoxA* gene expression at the distal edge of the developing limb bud, where the 5′ *HoxA* genes are most strongly expressed. By stage 36, limbs depleted of *HOTTIP* RNA showed notable shortening and bending of distal bony elements, including the radius, ulna and third digit (~20% length reduction for each compared to contralateral and stage-matched limbs treated with control virus, $P < 0.05$, Student's *t*-test, Fig. 2f). This phenotype resembled some of the defects in mice lacking *HoxA11* and *HoxA13* (refs 21–23). Together, these data indicate that *HOTTIP* RNA controls activation of distal *Hox* genes *in vivo*.

The broad impact of *HOTTIP* RNA on gene activation across the *HOXA* locus is reminiscent of the broad domains of chromatin modifications demarcating active and silent chromosomal domains[12]. 5C analysis of control and *HOTTIP*-depleted cells showed little change in higher order chromosomal structure, indicating that the chromosomal looping is pre-configured and upstream of gene expression

©2011 Macmillan Publishers Limited. All rights reserved

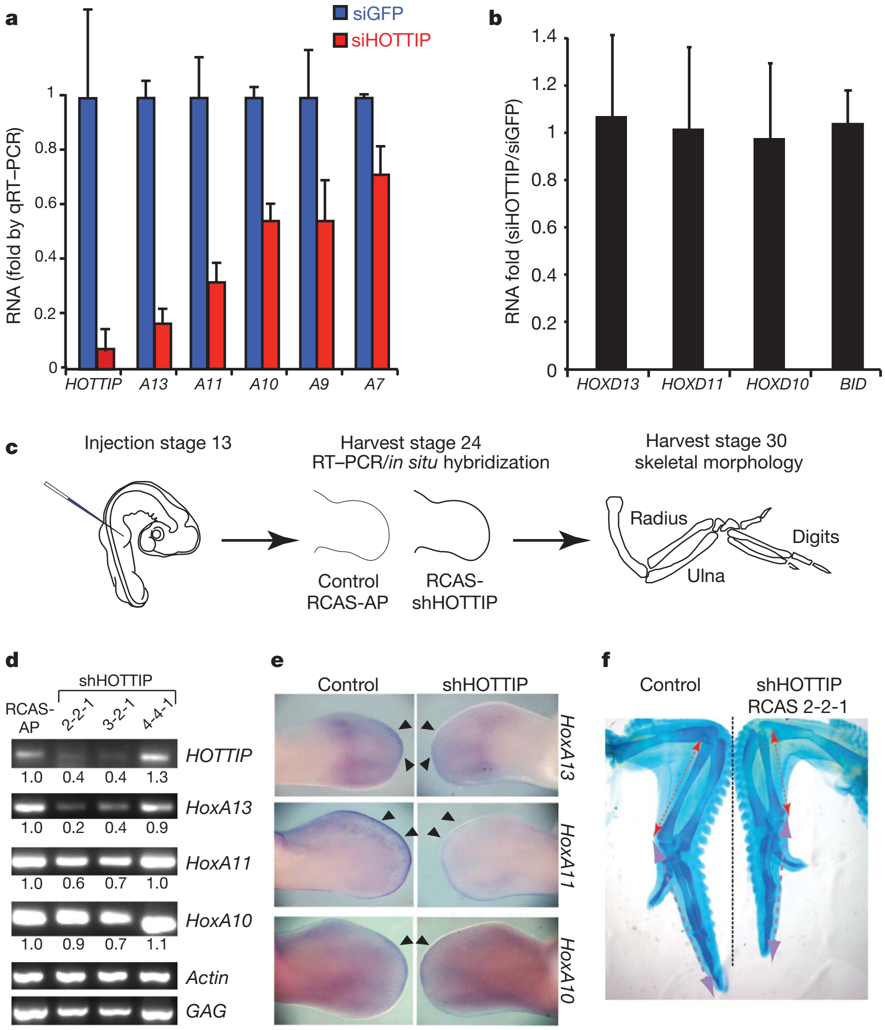


Figure 2 | ***HOTTIP* is required for coordinate activation of 5′ *HOXA* genes.** **a**, **b**, Knockdown of *HOTTIP* RNA abrogates expression of 5′ *HOXA* genes in foreskin fibroblasts (**a**), but not *HOXD* or *BID* genes (**b**). Means + s.d. are shown ($n = 3$). GFP, green fluorescent protein. **c**, Schematic of chick RNAi experiment. **d**, *HOTTIP* RNA is required for 5′ *HoxA* gene expression *in vivo*. RT–PCR of the indicated genes from control or *HOTTIP*-depleted distal limb bud is shown; quantification and normalization by *Actin* signal is shown below each band. *GAG* signal confirms successful retroviral transduction in all cases. **e**, *In situ* hybridization of 5′ *HoxA* genes in chick limb buds. Arrowheads highlight distal domains of high *HoxA* gene expression that are affected by *HOTTIP* knockdown. **f**, Shortening of distal bony elements in *HOTTIP*-depleted forelimbs. Alcian blue staining highlights the skeletal elements. Red and purple lines highlight radius and 3rd digit lengths, respectively.

(Supplementary Fig. 6a). In contrast, *HOTTIP* RNA knockdown led to broad loss of H3K4me3 and H3K4me2 across the *HOXA* locus, most prominently over 5′ *HOXA* and *HOTTIP* gene itself (Fig. 3a, Supplementary Figs 6b and 7). *HOTTIP* RNA knockdown also increased H3K27me3 focally over *HOTTIP* gene, but had little impact on H3K27me3 across *HOXA*. These results indicate that *HOTTIP* RNA is required for maintenance of H3K4me3 across the *HOXA*. These findings also imply that loss of 5′ *HOXA* gene transcription upon *HOTTIP* RNA knockdown is likely to be due to loss of H3K4me3 (or other changes) rather than ectopic spread of H3K27me3.

H3K4 methylation of the *HOX* loci is carried out by the MLL family of complexes[24]. In mammals, at least six MLL family members of SET-domain-containing lysine methyltransferases interact with a core complex of WDR5, ASH2L, RBBP5, as well as with other proteins, for substrate recognition and genomic targeting[24]. Genetic analyses indicate that MLL1 and 2 are most essential for *HOX* gene expression in fibroblasts[25], and MLL1 in particular is recruited to promoters of *HOX* genes to maintain their activation states[26]. In distally-derived human fibroblasts, MLL1 and WDR5 densely occupied extended region of the 5′ *HOXA* cluster, coincident with the H3K4me3 domain, with specific 'peaks' of occupancy near the transcriptional start sites (TSS) of multiple 5′ *HOXA* genes (Fig. 3b). Strikingly, *HOTTIP* RNA knockdown abrogated the peaks of MLL1 and WDR5 occupancy near TSS, resulting in diffuse and less intense binding of MLL1 and WDR5 across *HOXA* cluster, most prominently over the 5′ *HOXA* domain. *HOTTIP* RNA knockdown also led to increased accumulation of MLL1 and WDR5 on *HOTTIP* gene itself (Supplementary Fig. 8). Thus, *HOTTIP* RNA seems critical for maintaining a specific pattern of MLL complex occupancy across the *HOXA* locus to facilitate H3K4me3 and active transcription.

To define the molecular link between *HOTTIP* RNA and MLL complex, we reasoned that *HOTTIP* RNA may physically interact with one or more subunits of the MLL complex. Purified, *in-vitro*-transcribed, full-length *HOTTIP* RNA bound specifically to recombinant glutathione-*S*-transferase-conjugated WDR5 (GST–WDR5), but not to GST, RBBP5, ASH2L, or the telomeric protein TRF1 (also known as TERF1; Fig. 4a, b). The C terminus of MLL1, containing the SET domain, bound non-specifically to all RNAs, consistent with previous studies[27]. Immunoprecipitation of endogenous WDR5 from two different cell lines each specifically retrieved endogenous *HOTTIP* RNA (Fig. 4c), indicating that WDR5 and *HOTTIP* RNA interact in living

©2011 Macmillan Publishers Limited. All rights reserved

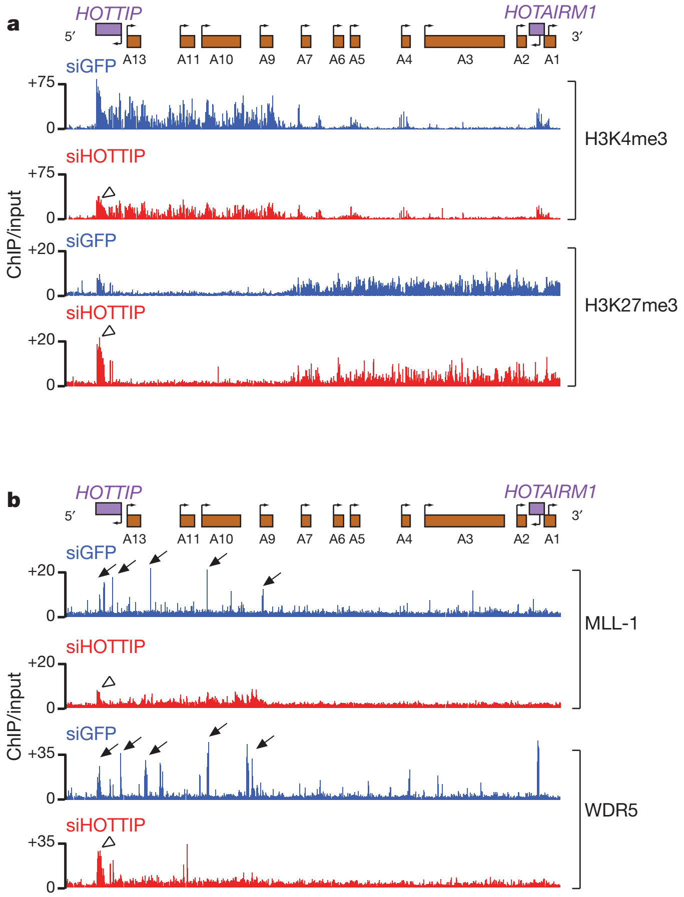


Figure 3 | *HOTTIP* RNA is required for the active chromatin state of 5′ *HOXA* cluster. **a**, Knockdown of *HOTTIP* RNA broadly decreases H3K4me3 across 5′ *HOXA* locus but focally affects H3K27me3 at *HOTTIP* gene. Display is as in Fig. 1a. **b**, Knockdown of *HOTTIP* RNA abrogates peaks of MLL1 and WDR5 occupancy near TSSs of 5′ *HOXA* genes and leads to accumulation of these proteins at *HOTTIP* gene itself. Arrows highlight peaks of MLL1 and WDR5 occupancy; open arrowheads highlight chromatin state of *HOTTIP* gene upon *HOTTIP* RNA knockdown.

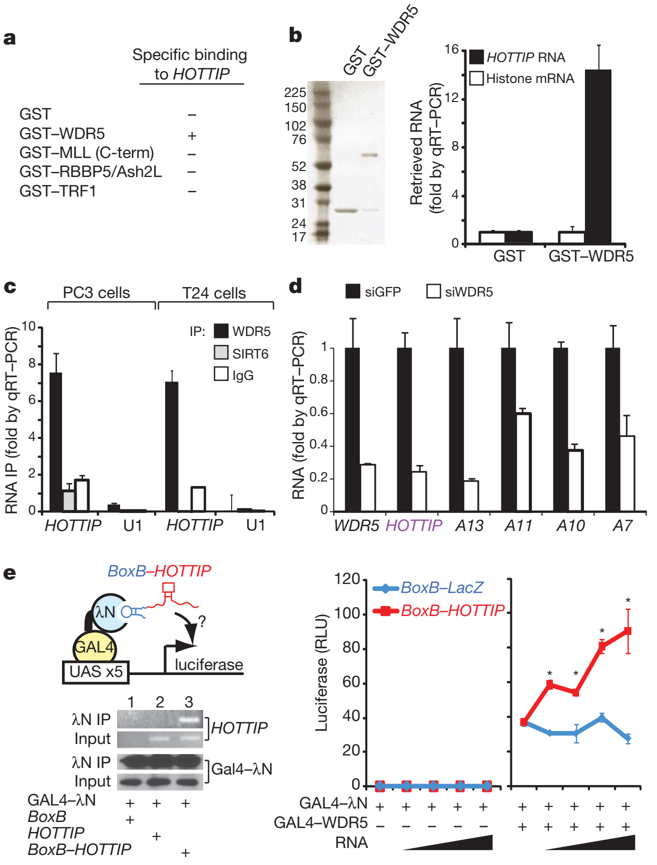


Figure 4 | *HOTTIP* RNA programs active chromatin via WDR5. **a**, Summary of RNA–protein interaction studies. Each of the indicated recombinant protein was purified and used to retrieve purified *HOTTIP* RNA or control histone RNA *in vitro*. Only GST–WDR5 specifically retrieved *HOTTIP*. **b**, *HOTTIP* RNA binds directly and specifically to WDR5. Left, purified GST and GST–WDR5 are visualized by SDS–PAGE and Coomassie Blue staining. Right, retrieved RNAs are quantified by qRT–PCR. **c**, *HOTTIP* RNA binds specifically to WDR5 in cells. Immunoprecipitation (IP) of endogenous WDR5 protein from PC3 (prostate) and T24 (bladder) carcinoma cells specifically retrieved *HOTTIP*, but not control IPs with IgG or chromatin binder SIRT6. U1 spliceosomal RNA served as negative control. **d**, WDR5 is required for 5′ *HOXA* gene expression, including *HOTTIP* RNA. **e**, *HOTTIP* RNA recruitment potentiates transcription. Left, the *BoxB* tethering system. *BoxB*–RNA specifically binds λN fused to GAL4 DNA binding domain, recruiting the complex to a UAS-luciferase reporter gene. After transient transfection, IP of GAL4- λN specifically retrieves *BoxB–HOTTIP* RNA. Right, luciferase activity after co-transfection of the indicated constructs. $*P < 0.05$ Student's *t*-test comparing *BoxB–LacZ* versus *BoxB–HOTTIP*). Sloped triangle indicates increasing input of plasmids encoding ncRNAs. Means ± s.d. ($n = 3$) are shown for all panels.

cells. Immunoprecipitation of an epitope-tagged WDR5 from a stable cell line that previously enabled stoichiometric purification of WDR5-interacting proteins[28] also specifically retrieved *HOTTIP* RNA (Supplementary Fig. 9). Knockdown of WDR5 broadly inhibited expression of 5′ *HOXA* genes, and also abrogated *HOTTIP* transcription, demonstrating mutual interdependence between *HOTTIP* RNA and WDR5 (Fig. 4d).

HOTTIP RNA seems to regulate genes *in cis*, due to its low copy number, distance dependence of *HOXA* target gene activation on endogenous *HOTTIP*, and the physical proximity of *HOTTIP* and its target genes as seen in 5C. Indeed, ectopic expression of *HOTTIP* RNA by retroviral transduction of lung fibroblasts, which do not express *HOTTIP*, failed to activate expression of distal *HOXA* genes, and did not change H3K4me3 and H3K27me3 patterns across *HOXA* (Supplementary Fig 10). Moreover, in foreskin fibroblasts that express endogenous *HOTTIP*, ectopic *HOTTIP* expression did not induce 5′ *HOXA* genes, nor rescue the effects of depleting endogenous nascent *HOTTIP* RNA (Supplementary Fig. 11). The lack of response in foreskin fibroblasts is notable because endogenous *HOTTIP* RNA is active in these cells, indicating that the protein partners of *HOTTIP* are all present and target genes are receptive. Ectopically expressed *HOTTIP* RNA, being transcribed from retroviral insertion sites scattered randomly in the genome, may not be able to find 5′*HOXA* genes. In contrast, endogenous *HOTTIP* RNA is directly positioned near the 5′ *HOXA* genes by chromosomal looping, allowing interaction and control.

To test the requirement of an exogenous targeting mechanism, we engineered an allele of *HOTTIP* RNA that can be artificially recruited to a reporter gene. Addition of five copies of the *BoxB* RNA element[29] to *HOTTIP* RNA allows the fusion transcript to be recruited to the λN RNA binding domain fused to a GAL4 DNA-binding domain (Fig. 4e). Recruitment of *HOTTIP* RNA to a silent GAL4 promoter is not sufficient to initiate transcription, but can significantly boost transcription if the promoter is also bound by WDR5 and transcriptionally active (Fig. 4e). By uncoupling the sites of *HOTTIP* transcription versus *HOTTIP* RNA function, this experiment indicates that the proximity of *HOTTIP* RNA—rather than the act of transcription—maintains target gene expression. To further support the functionality of *HOTTIP* RNA, deletion analysis identified a ~1 kb domain in the 5′ of *HOTTIP* RNA (*HOTTIP*$^{\text{Exons 1–2}}$) that retains WDR5 binding activity (Supplementary Fig. 12a). Enforced overexpression of *HOTTIP*$^{\text{Exons 1–2}}$

©2011 Macmillan Publishers Limited. All rights reserved

in foreskin fibroblasts inhibited 5′ *HOXA* gene expression in an apparently dominant negative manner (Supplementary Fig. 12b).

In summary, *HOTTIP* RNA is a key locus control element of *HOXA* genes and distal identity. Chromosomal looping brings *HOTTIP* RNA in close proximity to the 5′ *HOXA* genes. *HOTTIP* transcription acts as a switch to produce *HOTTIP* lincRNA, which binds to and targets WDR5–MLL complexes to the 5′ *HOXA* locus, yielding a broad domain of H3K4me3 and transcription activation (Supplementary Fig. 13). The mutual interdependence between *HOTTIP* RNA and WDR5 creates a positive feedback loop that maintains the ON state of the locus. These findings provide an integrated view linking three dimensional genome organization to dynamic programming of chromatin states, and ultimately to developmental pattern formation.

H3K4 methylation is a feature of almost all transcribed genes, and MLL family proteins are involved in many cell fate decisions in development and disease[24]. Our findings suggest that additional lincRNAs, especially those associated with enhancers or enhancer-like activities[8–10], may also be involved in gene activation by programming active chromatin states, and highlight WDR5 and other WD40 repeat proteins as candidate adaptors that link chromatin remodelling complexes to lincRNAs. *Cis*-restricted lincRNAs may be ideally suited to link chromosome structure and gene expression. Because such lincRNA can only act on its neighbours in space, information in higher order chromosomal looping can be faithfully transmitted to chromatin modification via RNA recruitment of enzymatic activities, and thus into gene expression.

METHODS SUMMARY

High throughput chromosome confirmation capture (5C) was performed on foreskin and lung fibroblasts, as well as foreskin fibroblasts treated with control or siRNA against *HOTTIP* RNA, as described[13]. siRNA knockdown experiments on cultured human fibroblasts and qPCR were performed as described previously[12]. ChIP-chip was performed as described[12] using ultra-high-density *HOX* tiling arrays. Full-length *HOTTIP* RNA was cloned by 5′ and 3′ rapid amplification of cloned/cDNA ends (RACE). Single-molecule RNA-fluorescent in situ hybridization (FISH) was performed using a pool of fluorescently-labelled oligonucleotides specific to *HOTTIP* RNA. *In vivo HOTTIP* RNA knockdown in chick was accomplished by microinjecting retroviruses carrying shRNA into prospective wing and leg buds and the animals were harvested at 2, 4 and 9 weeks post injection for RNA *in situ* hybridization, immunohistochemistry and whole-mount limb analysis, respectively. RNA-immunoprecipitation with WDR5 was performed as described[12]. Tethering experiments were done in 293T cells with co-transfections of various constructs containing a upstream activating sequence (UAS)-luciferase reporter, GAL4-WDR5, *BoxB* alone, and *BoxB* fused to full-length *HOTTIP* or full-length *LacZ*; cells were lysed 48 h after transfection and luciferase activity was determined.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 February 2010; accepted 12 January 2011.

Published online 20 March; corrected 7 April 2011 (see full-text HTML version for details).

1. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nature Rev. Genet.* **10,** 155–159 (2009).
2. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136,** 629–641 (2009).
3. Sanchez-Elsner, T., Gou, D., Kremmer, E. & Sauer, F. Noncoding RNAs of trithorax response elements recruit *Drosophila* Ash1 to *Ultrabithorax*. *Science* **311,** 1118–1123 (2006).
4. Petruk, S. *et al.* Transcription of *bxd* noncoding RNAs promoted by trithorax represses *Ubx* in *cis* by transcriptional interference. *Cell* **127,** 1209–1221 (2006).
5. Dinger, M. E. *et al.* Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* **18,** 1433–1445 (2008).
6. Dean, A. On a chromosome far, far away: LCRs and gene expression. *Trends Genet.* **22,** 38–45 (2006).
7. Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P. & Proudfoot, N. J. Intergenic transcription and transinduction of the human β-globin locus. *Genes Dev.* **11,** 2494–2509 (1997).
8. De Santa, F. *et al.* A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biol.* **8,** e1000384 (2010).
9. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).
10. Ørom, U. A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143,** 46–58 (2010).
11. Chang, H. Y. Anatomic demarcation of cells: genes to patterns. *Science* **326,** 1206–1207 (2009).
12. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human *HOX* loci by noncoding RNAs. *Cell* **129,** 1311–1323 (2007).
13. Dostie, J. *et al.* Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16,** 1299–1309 (2006).
14. Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B. & Cavalli, G. Genome regulation by polycomb and trithorax proteins. *Cell* **128,** 735–745 (2007).
15. Zhang, X. *et al.* A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. *Blood* **113,** 2526–2534 (2009).
16. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).
17. Kmita, M., Fraudeau, N., Herault, Y. & Duboule, D. Serial deletions and duplications suggest a mechanism for the collinearity of *Hoxd* genes in limbs. *Nature* **420,** 145–150 (2002).
18. Harpavat, S. & Cepko, C. L. RCAS-RNAi: a loss-of-function method for the developing chick retina. *BMC Dev. Biol.* **6,** 2 (2006).
19. Nelson, C. E. *et al.* Analysis of *Hox* gene expression in the chick limb bud. *Development* **122,** 1449–1466 (1996).
20. Kmita, M. *et al.* Early developmental arrest of mammalian limbs lacking *HoxA/HoxD* gene function. *Nature* **435,** 1113–1116 (2005).
21. Small, K. M. & Potter, S. S. Homeotic transformations and limb defects in Hox A11 mutant mice. *Genes Dev.* **7,** 2318–2328 (1993).
22. Davis, A. P., Witte, D. P., Hsieh-Li, H. M., Potter, S. S. & Capecchi, M. R. Absence of radius and ulna in mice lacking *hoxa-11* and *hoxd-11*. *Nature* **375,** 791–795 (1995).
23. Fromental-Ramain, C. *et al. Hoxa-13* and *Hoxd-13* play a crucial role in the patterning of the limb autopod. *Development* **122,** 2997–3011 (1996).
24. Ruthenburg, A. J., Allis, C. D. & Wysocka, J. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Mol. Cell* **25,** 15–30 (2007).
25. Wang, P. *et al.* Global analysis of H3K4 methylation defines MLL family member targets and points to a role for MLL1-mediated H3K4 methylation in the regulation of transcriptional initiation by RNA polymerase II. *Mol. Cell. Biol.* **29,** 6074–6085 (2009).
26. Guenther, M. G. *et al.* Global and Hox-specific roles for the MLL1 methyltransferase. *Proc. Natl Acad. Sci. USA* **102,** 8603–8608 (2005).
27. Krajewski, W. A., Nakamura, T., Mazo, A. & Canaani, E. A motif within SET-domain proteins binds single-stranded nucleic acids and transcribed and supercoiled DNAs and can interfere with assembly of nucleosomes. *Mol. Cell. Biol.* **25,** 1891–1899 (2005).
28. Wysocka, J. *et al.* WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell* **121,** 859–872 (2005).
29. Baron-Benhamou, J., Gehring, N. H., Kulozik, A. E. & Hentze, M. W. Using the λN peptide to tether proteins to RNAs. *Methods Mol. Biol.* **257,** 135–154 (2004).
30. Lajoie, B. R., van Berkum, N. L., Sanyal, A. & Dekker, J. My5C: web tools for chromosome conformation capture studies. *Nature Methods* **6,** 690–691 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Tabin for chick *Hox* gene probes, M. Scott and members of our labs for input, and M. Lin for use of the confocal microscope and imaging expertise. Supported by grants from the California Institute for Regenerative Medicine (H.Y.C., J.W.), the National Institutes of Health (HG003143 to J.D.), and the Scleroderma Research Foundation (H.Y.C.). K.C.W. is a recipient of a Dermatology Foundation Career Development Award. J.D. is a recipient of the W. M. Keck Foundation Distinguished Young Scholar Award. H.Y.C. and M.L. are Early Career Scientists of the Howard Hughes Medical Institute.

Author Contributions K.C.W., R.A.G. and H.Y.C. initiated the project; K.C.W. and H.Y.C. designed the experiments; K.C.W., Y.W.Y., B.L., A.S., R.C.-Z., B.R.L., A.P., R.A.F., J.D. and J.A.H. conducted the experiments and analysed the data; Y.C. and M.L. purified the recombinant proteins; J.W. provided antibodies and cell lines; K.C.W. and H.Y.C. prepared the manuscript with inputs from all co-authors.

Author Information Sequence for human *HOTTIP* RNA has been deposited with GenBank under the accession number GU724873. Microarray data are deposited in Gene Expression Omnibus (GEO) under accession number GSE26540. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to H.Y.C. (howchang@stanford.edu).

©2011 Macmillan Publishers Limited. All rights reserved

METHODS

Cells. Primary human fibroblasts derived from different anatomic sites were as described[12,31–36]. Primary human fibroblasts in culture retain their positional identity and have been used to examine chromatin states associated with positional memory, which have been confirmed *in vivo*[33,34,37].

Chromatin immunoprecipitation followed by microarray analysis. ChIP-chip was performed using anti-H3K27me3 (Abcam), anti-H3K4me3 (Abcam), anti-H3K4me2 (Abcam), anti-histone H3 (Abcam), anti-PolII (Abcam), anti-MLL1 (gift of R. Roeder), and anti-WDR5[28] antibodies as previously described[12]. Chromatin from each indicated cell type or RNAi treatment is split into multiple tubes and subject to ChIP with different antibodies in parallel. Retrieved DNA and input chromatin were competitively hybridized to custom tiling arrays interrogating human *HOX* loci at 5-bp resolution as previously described[12].

5C analysis of the ENm010 HoxA1 region. 5C primers were designed at HindIII restriction sites using 5C primer design tools previously developed[13] and made available online at http://my5C.umassmed.edu (ref. 30). Reverse primers were designed for fragments overlapping a known transcription start site from GENCODE transcripts[38], or overlapping a start site as experimentally determined by CAGE Tag data of the ENCODE pilot project[39]. Forward primers were designed for all other HindIII restriction fragments. Primers were excluded if highly repetitive sequences prevented the design of a sufficiently unique 5C primer. Primers settings were: U-BLAST: 3; S-BLAST: 130: 15-MER: 1320; MIN_FSIZE: 40; MAX_FSIZE: 50000; OPT_TM: 65; OPT_PSIZE: 40. DNA sequence of the universal tails of forward primers was CCTCTCTATGGGCAGTCGGTGAT; DNA sequence for the universal tails of reverse primers was AGAGAATGAGGAACC CGGGGCAG. A 6-base barcode was included between the specific part of the primers and the universal tail. In total 17 reverse primers and 90 forward primers were designed in the 500 kb *HoxA1* locus (ENm010) and hence a total of 1,530 *cis* interaction were interrogated in this region. Primer sequences are available separately (Supplementary Table 1).

3C was performed with HindIII as previously described[40] separately for fetal lung and foreskin fibroblasts (FB) and also for the control and *HOTTIP* knockdown foreskin FBs. For the 5C reaction, a total of 107 forward and reverse primers of HoxA1 region were mixed with either the ENCODE random region (ENr) primer pool comprising of 2,673 forward and 523 reverse primers (covering 30 additional ENCODE regions) or the ENr313 primer pool comprising of 57 forward and 58 reverse primers (covering 1 additional ENCODE region). 5C was then performed in 10 reactions each containing an amount of 3C library that represents 200,000 genome equivalents and 1 fmol of each primer. The 5C analysis of *HoxA1* region was carried out in two biological replicates of fetal lung and foreskin FBs. 5C ligation products were amplified using a pair of universal primers that recognize the common tails of the 5C forward and reverse primers described above and pooled together. To facilitate paired-end DNA sequence analysis on the Illumina GA2 platform, paired-end adaptor oligonucleotides were ligated to the 5C library using the Illumina PE protocol and PCR amplification of the library was carried out for 18 cycles with Illumina PCR primer PE 1.0 and 2.0. The 5C library was then sequenced on the Illumina GA2 platform generating 36 base paired end reads. For fetal lung FBs we obtained 7,625,276 and 10,947,424 mapped reads for two biological replicates of which 1,339,861 and 242,301 could be specifically mapped back to interactions within ENm010 using Novoalign (http://www.novocraft.com), respectively. For two biological replicates of foreskin FBs we obtained 7,311,386 and 5,731,107 mapped reads of which 2,752,789 and 66,769 could be mapped back to the ENm010 region, respectively. In the case of the knockdown study, control green fluorescent protein (GFP) knockdown foreskin FB 5C library yielded 4,909,482 mapped reads whereas *HOTTIP* knockdown foreskin FB had 5,565,389 mapped reads of which 39,168 and 38,950 could be mapped back to ENm010 for control GFP and *HOTTIP* knockdown, respectively. In the set with fetal lung and foreskin fibroblast samples, 5C for ENm010 was multiplexed for deep sequencing with 5C of one other region, ENr313; in the set containing the knockdown samples, ENm010 was multiplexed with 5C of 30 other genomic regions. The different extent of multiplexing resulted in different number of sequencing reads mapping back to ENm010. In all instances the mappable reads were proportional to the degree of multiplexing, indicating equivalent library quality despite different read numbers. Supplementary Table 2 outlines the library composition of each experiment. The heat maps are scaled as follows—for Fig. 1a, distal (foreskin) FBs: 262–17,467, proximal (lung) FBs: 7–5,846; for Supplementary Fig. 6, siGFP: 1–100, siHOTTIP: 1–100. Raw data from the 5C experiments used to generate the binned heat maps in Fig. 1a and Supplementary Fig. 6 can be found in Supplementary File 1. Raw data are available by request.

***HOTTIP* cloning, sequence and expression analysis.** We previously identified a portion of *HOTTIP* as a non protein-coding transcribed region named ncHOXA13-96 (ref. 12). This region also overlaps expressed sequence tag (EST) clone AK093987 that was previously observed to be expressed in cancer cell lines derived from posterior anatomic sites[41]. 5′ and 3′ RACE (RLM Race kit, Applied Biosystems/Ambion) showed full-length *HOTTIP* RNA to be 3,764 nucleotides, extending the known transcribed region by more than 1,400 bases. BLAST and BLAT confirmed that portions of *HOTTIP* are well conserved in mammals and even in avians but had no protein coding potential. Full-length *HOTTIP* RNA sequence has been deposited at NCBI (accession number GU724873). qRT–PCR with SYBR Green was conducted as recommended by the manufacturer (Agilent Technologies). Primer sequences specific for *HOTTIP* were CCTAAAGCCACGC TTCTTTG (HOTTIP-F) and TGCAGGCTGGAGATCCTACT (HOTTIP-R). For Supplementary Fig. 11, endogenous nascent *HOTTIP* was distinguished from ectopic *HOTTIP* expressed from cDNA using primers that spanned intron–exon junctions.

Strand-specific RT–PCR. RNA extracted from primary foreskin fibroblasts was reverse transcribed (SuperScript III, Invitrogen) using combinations of the previously described *HOTTIP*-specific primers HOTTIP-F and/or HOTTIP-R as diagrammed in Supplementary Fig. 1b. Resulting cDNA was then PCR-amplified using both HOTTIP-F and HOTTIP-R primers to visually determine strand specificity.

***HOTTIP* transcript count per cell.** The level of *HOTTIP* transcript per cell was calculated from the level of *HOTTIP* in 500,000 cells. Full-length *HOTTIP* in pcDNA3.1+ was assayed by qPCR using primers HOTTIP-F and HOTTIP-R at predetermined concentrations in triplicate to generate a linear amplification curve dependent on the moles of template DNA (Supplementary Fig. 2). The qRT–PCR value from 500,000 foreskin fibroblasts was determined and plotted, and the corresponding total molecules of transcript was divided by 500,000 to determine the approximate number of transcripts per cell.

Single-molecule RNA fluorescence *in situ* hybridization (RNA-FISH). Single molecule RNA-FISH was performed as described in ref. 42 with the following modifications: the amount of hybridization solution per chamber was doubled to allow for proper coating of the chamber and the amount of glucose-oxidase buffer was tripled to assist in image acquisition. Images were acquired using an Olympus FV1000 confocal microscope within 2 h of the addition of the glucose-oxidase buffer.

RNA interference. Primary foreskin fibroblasts were transfected with siRNAs targeting *HOTTIP* and *WDR5* using Lipofectamine 2000 (Invitrogen) as per manufacturer's instructions. Total RNA was harvested 48–72 h later using TRIzol (Invitrogen) and RNeasy Mini Kits (Qiagen) as previously described[34]. For the intronic *HOTTIP* knockdown experiment in Supplementary Figure 11, a pool of 10 siRNAs (Supplementary Table 3) targeting intronic regions in *HOTTIP* were transfected into foreskin fibroblasts, and RNA isolated as above.

Generation of shRNAs against chicken *HOTTIP*. A reporter construct encoding a GFP–chicken *HOTTIP* fusion transcript was used in a small-scale screen to identify highly effective shRNA constructs. Eleven shRNAs targeting conserved regions of chicken *HOTTIP* were designed and inserted into the pSMP system (Thermo/Open Biosystems). The reporter construct and shRNA constructs were cotransfected into Phoenix cells, and *HOTTIP* transcript levels were analysed via reduced GFP fluorescence and by qRT–PCR. Three shRNAs that were effective *in vitro* were then cloned into RCAS vector for studies in chick embryos[18].

Chick RNAi. RCAS *HOTTIP* hairpin and RCAS AP viruses were made by transfecting DF-1 cells with viral DNA. Transfected DF-1 cells were grown and passaged, after which the virus-containing supernatant was collected, concentrated and titred. Fertilized chicken eggs were incubated in a humidified rotating incubator at 37 °C until they reached Hamilton/Hamburger stage 10. Eggs were then windowed to expose the embryos. After gently removing the vitelline membrane, chicken embryos were microinjected with RCAS-*HOTTIP* hairpins and RCAS-AP viruses at the prospective wing and leg buds. All viral stocks have titres of 1×10^8 IU ml^{-1}, and each limb was injected five times. The infected embryos were allowed to incubate at 37 °C and were harvested 2 or 4 days after injection to detect viral infection by immunohistochemistry. Total RNA was extracted from injected forelimbs, and RT–PCR analysis was performed 4 days after injection. Chicken embryos were harvested 9 days post-injection to carry out whole-mount Alcian blue staining. A total of 50 animals were injected.

Hairpin sequences for chick *HOTTIP* were TGCTGTTGACAGTGAGCGAC CCGAAGATGTGTCTGATTTGTAGTGAAGCCACAGATGTACAAATCAGA CACATCTTCGGGCTGCCTACTGCCTCGGA (2-2-1), TGCTGTTGACAGTG AGCGCCGCTCTGCTCTCCTCTCTCTCTAGTGAAGCCACAGATGTAGAG AGAGAGGAGAGCAGAGCGATGCCTACTGCCTCGGA (3-2-1), and TGCT GTTGACAGTGAGCGAATCCTTAATCGAATCTGATTTAGTGAAGCCACA GATGTAAAATCAGATTCGATTAAGGATCTGCCTACTGCCTCGGA (4-4-1).

***HOTTIP* overexpression.** Full-length *HOTTIP* and a truncated transcript consisting of exons 1 and 2 (*HOTTIP*$^{\mathrm{Exons\ 1-2}}$) were cloned into the LZRS vector (gift of P. Khavari), and then transfected into Phoenix cells (gift of G. Nolan) to generate amphotropic retroviruses. Primary human fibroblasts were infected with either LZRS-full length *HOTTIP* (lung), LZRS-truncated *HOTTIP* (foreskin), or LZRS-GFP (both lung and foreskin), then passaged over 60 days, with periodic

©2011 Macmillan Publishers Limited. All rights reserved

testing of *HOXA* and *HOTTIP* expression by qRT–PCR. These cells were used in the rescue experiments depicted in Supplementary Fig. 11.

GST pull-down. Full-length *HOTTIP*, truncated *HOTTIP* containing exons 1 and 2 (*HOTTIP*$^{\text{Exons 1–2}}$), and histone H2B1 mRNA were transcribed *in vitro* using T7 polymerase according to manufacturer's instructions (Promega), denatured, and refolded in folding buffer (100 mM KCl, 10 mM $MgCl_2$, Tris pH 7.0). GST-tagged WDR5, C-terminal MLL1, RBBP5/Ash2L and TRF1 were expressed in *Escherichia coli* and purified as described[43]. Each GST-fusion protein was bound to glutathione beads (Amersham/GE Healthcare) and blocked with excess yeast total mRNA in PB100 buffer (20 mM HEPES pH 7.6, 100 mM KCl, 0.05% NP40, 1 mM DTT, 0.5 mM PMSF) for 1 h at room temperature. Beads were then incubated with either *in-vitro*-transcribed *HOTTIP* or histone H2B1 mRNA for 45 min at room temperature. After three washes in PB200 buffer (20 mM HEPES pH 7.6, 200 mM KCl, 0.05% NP40, 1 mM DTT, 0.5 mM PMSF), bound RNAs were extracted and analysed by qRT–PCR, as previously described.

RNA immunoprecipitation. HeLa-WDR5-Flag cells: 48 h after Lipofectamine 2000-mediated transfection of *HOTTIP* into HeLa WDR5-Flag cells (approximately 10^7 cells), total protein was extracted as previously described, with modifications[44]. Briefly, cells were resuspended in Buffer A (10 mM HEPES pH 7.5, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, 1.0 mM PMSF), lysed in 0.25% NP40, and fractionated by low speed centrifugation. The nuclear fraction was resuspended and lysed in Buffer C (20 mM HEPES pH 7.5, 10% glycerol, 0.42 M KCl, 4 mM $MgCl_2$, 0.5 mM DTT, 1.0 mM PMSF). Combined nuclear and cytoplasmic fractions were immunoprecipitated with mouse anti-Flag M2 monoclonal antibody (Sigma) or mouse IgG affixed to agarose beads (Sigma) for 3 to 4 h at 4 °C. Beads were washed four times with wash buffer (50 mM TrisCl pH 7.9, 10% glycerol, 100 mM KCl, 5 mM $MgCl_2$, 10 mM β-mercaptoethanol, 0.1% NP40). After elution using Flag peptide (Sigma), co-immunoprecipitated RNA was extracted and analysed by qRT–PCR.

Endogenous WDR5 and SIRT6 RIP: cellular fractions were isolated as above and incubated with the anti-WDR5 (ref. 31) or anti-Sirt6 (ab62739, Abcam) antibodies overnight at 4 °C. Samples were washed in wash buffer, and co-immunoprecipitated RNA was extracted and analysed by qRT–PCR.

RNA chromatography. Full-length *in-vitro*-transcribed *HOTTIP* RNA was conjugated to adipic acid dehydrazide agarose beads as described[45]. The complexed beads were incubated with whole cell lysates from Hela WDR5-Flag cells, washed, and bound proteins visualized by western blotting.

***BoxB* tethering assay.** 293T cells were grown to about 50% confluence in 6-well plates on the day of transfection. Using Lipofectamine 2000 (Invitrogen), a plasmid encoding a luciferase gene under the control of five tandem GAL4 UAS sites were co-transfected with plasmids encoding GAL4-WDR5, GAL4-λN (the 22 amino acid RNA-binding domain of the lambda bacteriophage antiterminator protein N) peptide fused to a C-terminal GFP tag, *BoxB* (containing five repeats of the λN-specific 19 nucleotide binding site), *BoxB* fused to full-length *LacZ*, or *BoxB* fused to full-length *HOTTIP*. Cells were lysed 48 h after transfection, and luciferase assay kit (Promega) was used to determine relative levels of the luciferase gene product, following the manufacturer's protocol.

31. Chang, H. Y. *et al.* Diversity, topographic differentiation, and positional memory in human fibroblasts. *Proc. Natl Acad. Sci. USA* **99,** 12877–12882 (2002).
32. Chang, H. Y. *et al.* Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol.* **2,** 206–214 (2004).
33. Bernstein, B. E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120,** 169–181 (2005).
34. Rinn, J. L., Bondre, C., Gladstone, H. B., Brown, P. O. & Chang, H. Y. Anatomic demarcation by positional variation in fibroblast gene expression programs. *PLoS Genet.* **2,** e119 (2006).
35. Rinn, J. L. *et al.* A dermal *HOX* transcriptional program regulates site-specific epidermal fate. *Genes Dev.* **22,** 303–307 (2008).
36. Rinn, J. L. *et al.* A systems biology approach to anatomic diversity of skin. *J. Invest. Dermatol.* **128,** 776–782 (2008).
37. Soshnikova, N. & Duboule, D. Epigenetic temporal control of mouse *Hox* genes in vivo. *Science* **324,** 1320–1323 (2009).
38. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7** (Suppl 1), S4 (2006).
39. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
40. Dostie, J. & Dekker, J. Mapping networks of physical interactions between genomic elements using 5C technology. *Nature Protocols* **2,** 988–1002 (2007).
41. Sasaki, Y. T., Sano, M., Kin, T., Asai, K. & Hirose, T. Coordinated expression of ncRNAs and *HOX* mRNAs in the human *HOXA* locus. *Biochem. Biophys. Res. Commun.* **357,** 724–730 (2007).
42. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods* **5,** 877–879 (2008).
43. Smith, D. B. & Johnson, K. S. Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione *S*-transferase. *Gene* **67,** 31–40 (1988).
44. Dignam, J. D., Lebovitz, R. M. & Roeder, R. G. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res.* **11,** 1475–1489 (1983).
45. Michlewski, G. & Caceres, J. F. RNase-assisted RNA chromatography. *RNA* 16, 1673–1678 (2010).

©2011 Macmillan Publishers Limited. All rights reserved

COLUMN How to overcome procrastination on your research project **p.127**

NATURE NETWORK For further discussions on employment issues, visit **go.nature.com/lm1x4t**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

R. CUMMINS/CORBIS

San Diego hosts a cluster of biotechnology start-ups and medical companies that offer opportunities to researchers brave enough to take the leap.

CALIFORNIA

Safe harbour

San Diego's diverse corporate science portfolio offers opportunities for open-minded scientists hoping to escape stagnation in academia.

BY KAREN KAPLAN

If he were living just about anywhere else, Peter Teriete would be forced to leave a city and a lifestyle he enjoys. Teriete wants to start his own lab after seven years of postdoc positions, but is finding it hard to secure a faculty role at Sanford-Burnham Medical Research Institute in La Jolla, California — part of the greater San Diego area — where he is currently working as a structural biochemistry postdoc. His principal investigator would like to promote him to staff scientist, but is struggling to find funding. And Teriete hits a wall every time he has an interview for an academic research job anywhere else. "They all say, 'The position is yours if you bring a grant. You're doing very interesting research, but we have no funding to start you up,'" he says.

Teriete's academic straits are not unusual. Tight budgets across the United States have forced countless young researchers into tough decisions and professional uncertainty. And like many universities, institutions in San Diego are struggling. Academic hiring is especially problematic for young researchers.

Most scientists would have to cast a wide net, looking for a promising position in another city. But his San Diego location helps Teriete's situation considerably: the region offers options in several sectors and fields. With a child on the way, Teriete and his fiancée want to remain where they are — and an industry job may be the answer. "I've made a mental crossover," he says. "For a long time, I was set on the academic track, but I've become much more open in terms of making the transition into industry."

San Diego has a strong biotechnology cluster, in which collaborations make start-ups feasible for the entrepreneurially minded. Venture capital is starting to rebound, and niche research areas such as medical diagnostics are showing strong signs of growth (see 'Medicine on the move'). In tough economic times, the region is a relative success — and scientists willing and able to take their research strengths to a new field or sector will find opportunities to shine.

FALSE START

Greater San Diego is home to more than 80 research institutes, which collectively bring in about US$1 billion a year in federal research funding — among them such heavy hitters as the University of California, San Diego (UCSD); Scripps Research Institute, a centre for biomedicine; the Salk Institute for Biological Studies; and Sanford-Burnham, all in La Jolla. But ▸

© 2011 Macmillan Publishers Limited. All rights reserved

hiring is slow or nonexistent, and institutions say that taking on junior faculty members without external funding is simply too expensive.

Scripps has only in the past six months been able to hire 14 early-career researchers as part of a $28-million, five-year grant from Novartis, a pharmaceutical company based in Basel, Switzerland. Under the grant's terms, Scripps appointed scientists working in areas of interest to both itself and Novartis; the company retains the right to license all technology that the researchers produce in their first five years at Scripps. The institute could not have hired without the grant, says James Williamson, a microbiologist and dean of graduate studies at Scripps. "The funding climate has crunched the institutional budget, making it difficult to set aside the necessary start-up package," he says.

The outlook for early-career faculty members is little better at Sanford-Burnham, where young applicants face long odds. The hiring of established, mid-career researchers requires less of a dip into the institution's endowment, as many come with external grants and don't require start-up funds. Postdocs, however, are the focus of ongoing recruiting efforts, says Guy Salvesen, director of scientific training at the institute; he estimates that Sanford-Burnham takes on 10–15 a year.

At UCSD, hiring for state-funded positions has slowed. Again, researchers with extramural funds are the most sought-after recruits, says Kim Barrett, dean of graduate studies. Still, she says, junior-faculty opportunities at UCSD are likely to materialize in the next few years, thanks to a clinical and translational medicine building slated for completion in 2016 and funded in part by a $37.2-million award last year from the US National Institutes of Health (NIH). UCSD will probably recruit specialists in biomedical informatics, electronic health-record technology, diagnostic imaging and telemedicine — the use of communications technology to deliver clinical care.

BIOTECH RESILIENCE

Industrial research is better off. The region hosts a few international pharmaceutical companies — including Illumina; Life Technologies in Carlsbad, part of Greater San Diego; Takeda Pharmaceuticals of Osaka, Japan; and Johnson & Johnson of New Brunswick, New Jersey — but nowhere near as many as other US areas, such as Boston, Massachusetts. San Diego's biomedical industry is instead composed mostly of small boutiques and spin-off companies, many of which are too new and small to hire large numbers of researchers.

"There are opportunities outside the traditional circles for highly trained scientists."

Still, between 2009 and 2010, the cluster — which includes some 600 small drug-makers, medical-device companies, lab-supply firms and medical-diagnostics manufacturers — added jobs, in contrast to the industry's performance almost everyhwhere else in the state, says a report published in February by the California Healthcare Institute in La Jolla and the London-based professional-services firm PricewaterhouseCoopers. About 140 contract-research organizations are also headquartered in San Diego. Venture-capital investment in the cluster, predictably, slowed in 2008 as the recession raged, but started to recover in 2009 and 2010, according to an online report by PricewaterhouseCoopers.

But as in other US cities, the venture capital available often remains insufficient to sustain a fledgling company. Sridhar Prasad, a structural biologist who launched drug-screening start-up CalAsia Pharmaceuticals in September 2009, still holds out hope that new-found capital will find its way to his business, but is pursuing other avenues. So far, Prasad — who was laid off from his group-leader position at another local biotech in January 2009, just before the company went belly-up, and who now employs three researchers besides himself — has been keeping his fledgling company afloat with grants from the Michael J. Fox Foundation in New York, and is awaiting the outcome of several NIH grant applications. In mid-April, he will apply for another, larger NIH grant. Meanwhile, he says, he will carry on — and keep scouting around for either an investor to infuse capital or a larger company that wants to buy up his business.

Despite the risks, start-ups look more attractive than academia to many. Anjuli Timmer and her husband, John, decided to head straight to industry after their postdocs at UCSD.

"My principal investigator was trying to get me to pursue a tenure-track position, but I didn't want to," says Anjuli Timmer. She knew that getting a faculty position would have been nearly impossible, so instead she applied to several local biotech firms, and got offers from two. She shortened her postdoc from three years to two to accept a job as a microbiologist at Tanabe Research Laboratories in La Jolla, and hasn't looked back. She enjoys studying biological therapeutics for inflammatory diseases. "This is a once-in-a-lifetime opportunity, to work at a start-up that has a rich parent," she says, referring to the biotech's acquisition by Mitsubishi Tanabe Pharma of Osaka last year. She also likes the good pay, which started at $82,000 a year.

John Timmer, a molecular pathologist, accepted a position at start-up biotech Inhibrx last June. He also interrupted his postdoc to take the job, for which he was recruited by a former lab colleague from his days as a PhD student. John Timmer likes working in a company with just five colleagues, where everyone does everything — from assays to antibody development. "It's been really exciting," he says.

Ultimately, entrepreneurialism is the driving force behind the San Diego region's scientific success, say Lynn Reaser, chief economist at the Fermanian Business & Economic Institute at Point Loma Nazarene University in San Diego, and Mary Walshok, an industrial sociologist at UCSD. The area offers satisfying work for early-career researchers who can become entrepreneurs themselves, like Prasad; embrace the risk of working at a start-up, like the Timmers; or step into an emerging area such as telemedicine. "There are opportunities outside the traditional circles for highly trained scientists," says Walshok. ■

Karen Kaplan *is assistant editor for Careers.*

OPPORTUNITIES

Medicine on the move

The San Diego region is well known for its biotechnology sector, but the area's best near-term opportunities could be in niche fields such as medical diagnostics and wireless health — the use of mobile devices such as smart phones and wearable sensors to facilitate quick, convenient transmission of health data. Mary Walshok, an industrial sociologist at the University of California, San Diego, who investigates the economic forces that affect job trends, says that the city's cluster of companies and institutions already hosts 1,100 information-technology companies, with 50 specializing in wireless health. Interdisciplinary skills in medicine, bioinformatics and information technology are key. "Maybe it's not where the traditional PhD wants to go, but it's where a lot of the opportunities are going to be — the interface between physiological systems and devices," says Walshok.

Companies will continue to need researchers who can design wireless systems and interpret, analyse and manage the colossal streams of data that they will generate on everything from blood pressure to medication dosage, says Don Casey, chief executive of the West Wireless Health Institute in San Diego, a private non-profit research organization established in 2009. He predicts major growth in the bioinformatics and biodata analysis fields. "San Diego is a great place to do all this because there's already a huge amount of telecommunications infrastructure and a legacy of medical infrastructure," says Casey. "The young industry needs people well skilled in understanding how to examine large quantities of medical data and turn that into usable information." K.K.

© 2011 Macmillan Publishers Limited. All rights reserved

COLUMN

Waiting for the motivation fairy

It's easy to give in to procrastination — but **Hugh Kearns** and **Maria Gardiner** offer some tips for getting your drive back.

"I love deadlines. I love the whooshing sound they make as they go by."

— Douglas Adams

If you were trying to set up ideal conditions for procrastination, conducting a research project would provide them. Such projects tend to be large and time-consuming: completing a doctoral research project, for example, often takes three years or more. Deadlines and endpoints are often fuzzy and ill-defined. Then there's the reward structure: you can put in a lot of effort with little to no positive feedback along the way, and the rewards, if there are any, take a long time to come. Add to this the fact that scientists are often perfectionists with demanding, if not idealistic, expectations, and it is little wonder that procrastination is the most discussed topic in our graduate-student and researcher workshops. Many researchers simply take for granted that they are at the mercy of the forces of procrastination, doomed to increased stress levels and stretched deadlines. But there are simple strategies for pushing yourself to get engaged. The first is to recognize the patterns that you're falling into.

L. LATULIPPE/CORBIS

ADVANCED DISPLACEMENT

Some procrastination activities are pretty obvious. There's the morning coffee break that creeps into lunchtime. Or watching videos on YouTube and sending them to all your friends. Or updating your Facebook status when you should be updating your lab book.

But most procrastination is far more subtle, and can even be mistaken for productive work. For example, you might try to track down that elusive reference, even though you've already got more than you will ever have time to read. Or you could start a new experiment instead of analysing the old one. Or take stock of the glassware in the lab. Or check your e-mail. These activities make it seem as though you're doing something useful, and you may well be, but it's not the thing you should be doing right now.

So why is housekeeping, for example, so much fun when you're supposed to be working on your dissertation or a paper? It's a displacement activity, used to dispel the self-reproach or discomfort that we feel for not doing something else. Reading a novel or taking a nap causes too much guilt. But have you ever, say, reorganized your folders to make it easier to find the files? It would speed up your writing, after all. Or perhaps you've diligently labelled all the cupboards in the lab to make it easier to find things.

Although these activities or excuses seem acceptable, their fatal flaw is that once they're over, you still haven't finished that article, started that experiment or written your dissertation. You probably have an increased sense of guilt because you're not making progress on your goal. And although you've found and read that reference, you still don't feel motivated to write. Sadly, while you were answering e-mails or counting the glassware, the motivation fairy didn't stop by and make that difficult task look any more appealing. That's just not how motivation works.

Most people have a fundamental misunderstanding: we like to think that motivation leads to action, or, more simply, that when you feel like doing something, you'll do it. This model might work for things you enjoy doing, such as watching a film or going for a walk. But it's not particularly good for huge tasks with fuzzy deadlines. The problem is that you may never feel motivated to revise and resubmit that paper — at least not until a hard-and-fast deadline appears. You need a different model.

MOTIVATION MOJO

Some psychology research shows that action leads to motivation, which in turn leads to more action. You have to start before you feel ready; then you'll feel more motivated, and then you'll take more action. You've probably had this experience yourself. You put off running an analysis for ages; eventually, you decide to do it, and once you start, you say to yourself, "This isn't as bad as I thought. Why not keep going while I'm at it?"

Of course, starting before you feel motivated is difficult. But certain strategies can directly tackle the conditions that lead to procrastination in the first place.

First, big projects need to be broken down into steps. Not just small steps, but tiny steps. Instead of saying you'll make the revisions to the paper — which probably seems overwhelming — the tiny step could be that you'll read the reviewer's comments or you'll make the first two changes. Second, you need to set a time or deadline by which to perform that tiny step. Saying you'll do it later or tomorrow isn't enough — the deadline needs to have an 'o'clock' attached to it. Third, you need to build in an immediate reward. If you finish reading the comments by your deadline at 10:00 a.m., you can allow yourself to have a coffee, a brief chat or a quick e-mail exchange. It's highly likely that once you start the task, your motivation will kick in and you'll find yourself wanting to spend longer at it.

So if the motivation fairy hasn't been stopping off at your lab or desk very frequently, perhaps you should give her a hand. The next time you catch yourself engaging in displacement activities, remember that there's a way to recover that elusive drive. Follow our three rules and watch your motivation grow. ■

Hugh Kearns *and* **Maria Gardiner** *lecture and conduct research in psychology at Flinders University in Adelaide, Australia, and run workshops for graduate students and advisers (see ithinkwell.com.au).*

© 2011 Macmillan Publishers Limited. All rights reserved

SHIFT

A structured revolution.

JACEY

BY LIZ WILLIAMS

Hanson did not want to watch the old man die, but a sense of social responsibility kept him by the bedside until quite late into the night. Robbins's breathing was laboured, but there had been little change for several hours and so Hanson stepped out onto the balcony of the hospital to get some air. Below, the lights of the city spread out in a lacy sprawl, the red light at the summit of Canary Wharf blinking through the rain. Hanson took a deep breath of dirty London damp and tried to straighten his thoughts. Stress, and a lack of sleep, were beginning to take their inevitable toll, and on top of that was anxiety about the future.

Robbins's death would throw everything into doubt. The future of the funding programme had depended on him: he'd done all the legwork for it, calling in contacts in government and the various research councils. It didn't matter that Hanson had serious doubts about the project; the department had to stick together and he was a big believer in team playing, especially given the latest round of cuts. He'd seen enough of academic backstabbing not to want to see the Physics Department go the same way, mired down into a bloody morass of petty squabbling. For all his faults — dictatorial, overbearing, not-listening-to-reason, stubborn — all the things that Hanson had called him over the years, Robbins had held the department together and Hanson was willing to respect that. And now Robbins was dying, felled by the tumour that had, Hanson had learned only yesterday, been growing in his lung for the past eight months.

He braced his hands against the rough wet concrete and looked down for a dizzying moment into the street below. If Robbins had only said something — but of course, he hadn't. He'd preferred it this way. Hanson suspected that the old man had really wanted to drop dead in the lab and be buried in his stained white coat, a workhorse to the end. And in spite of all his noble thoughts about keeping the department on an even keel, the fact remained: beneath the shock and the genuine sadness, there was a growing, guilty relief.

Because now they wouldn't be committed to the project.

Hanson was next in line for departmental head. Not a done deal, of course, but highly likely. He'd wanted it for a long time, but not at this kind of expense. Now, he wondered a little at his own naivety, given that Robbins had always expressed the intention not to retire until absolutely necessary. Hanson knew that neither Benjamin nor Eleanor was committed to Robbins's theories: like himself, they considered that line of enquiry to be old-fashioned. Ben had expressed enough doubts along the line and so had Eleanor.

"It's had its chips." Eleanor's voice echoed in his head. "When you look at what the Yanks are doing — we ought to be going all out into that research, not piddling about with outdated theories, never mind what the old guard thinks."

➲ NATURE.COM
Follow Futures on Facebook at:
go.nature.com/mtoodm

"Trouble is," Hanson remembered replying, "it's not outdated. Ever studied Kuhn, Eleanor?"

"I'm not much of a philosopher."

"He said that the paradigm couldn't change until the old guard died. Literally."

They'd laughed about it then, but now it seemed sad, and not funny. He thought about it, leaning on the wet stone and feeling the rain cold against his skin. Robbins with his insistence on the old paradigm, on structures, on strings … Numbers raced through his mind, unspooling like skeins of thread. It wasn't the answer, he was sure of it. He had the new experiments all planned out and now that there was a real chance of implementation, of testing, he was conscious of a growing burst of excitement, a keen anticipation …

There were raised voices in the room beyond, urgent. He raced back inside, in time to see the screen flatline. "He's gone," a nurse said. And Hanson turned back to the city beyond the windows. There was a line of light, far on the horizon; threads of brightness spreading through the rain-wet stone. It lasted only for a moment, then it, too, was gone, but he could feel the change inside himself. Not just the paradigm that shifts, when its last exponent dies — but as he tried to hold onto the thought, it ebbed away. ■

Liz Williams *has been published by Bantam Spectra in the United States and Tor Macmillan in the United Kingdom, and is currently on her 15th novel.*

© 2011 Macmillan Publishers Limited. All rights reserved

Isotope fractionation in silicate melts by thermal diffusion

ARISING FROM F. Huang, P. *et al. Nature* **464,** 396–400 (2010)

It was recently shown that relatively large (compared to analytical precision) steady state thermal isotope fractionations are produced in silicate melts whenever temperature differences are maintained for a sufficiently long time[1,2]. Huang *et al.*[3] reported new data on thermal isotopic fractionation of magnesium, calcium, and iron in silicate liquids, and claimed (1) that thermal isotopic fractionations in silicate liquids are independent of composition and temperature, and (2) that their "results lead to a simple and robust framework for characterizing isotope fractionations by thermal diffusion in natural and synthetic systems". Here I consider whether the data and arguments presented by Huang *et al.*[3] support their claims. In summary, I caution against assuming (on the basis of the data presented by Huang *et al.*[3]) that the thermal isotopic fractionations are independent of temperature and composition, or that a framework of the type claimed has been found.

Huang *et al.*[3] reported new data on thermal isotopic fractionation of magnesium, calcium and iron in silicate liquids that extend the earlier results[1,2] to a larger temperature range and to compositions other than basalt. The magnesium isotopic composition versus temperature reported by Huang *et al.*[3] for molten andesite and basalt are shown in Fig. 1. The local slope in such a plot determines the thermal-diffusion isotopic sensitivity Ω, defined as the magnitude of isotopic fractionation per temperature offset, and expressed in units of ‰ °C^{-1} AMU^{-1} (here AMU indicates atomic mass unit). If Ω is independent of temperature, the data would fall on straight lines, which is clearly not the case for the andesite experiments (squares in Fig. 1), where the slope of the isotopic fractionations versus temperature shows that Ω_{Mg} varies from about 0.05 ‰ °C^{-1} AMU^{-1} for $T > 1{,}450$ °C to about 0.02 ‰ °C^{-1} AMU^{-1} for $T < 1{,}450$ °C. Despite this, Huang *et al.*[3] imply in their figure 1 that $\Omega_{\mathrm{Mg}} = 0.03$ ‰ °C^{-1} AMU^{-1} with an uncertainty of only ±10%. The iron isotope data reported by Huang *et al.*[3] also show large variations of Ω_{Fe} with temperature.

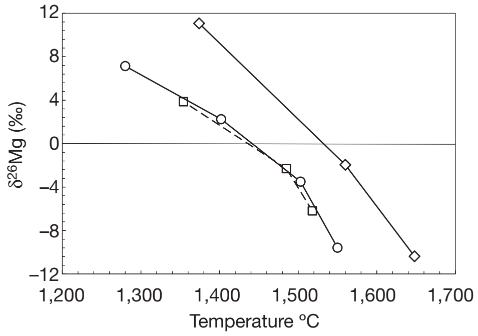


Figure 1 | Isotopic fractionation of magnesium as a function of temperature. Data are taken from table 1 in Huang *et al.*[3]. Shown are values for molten Mount Hood andesite (squares for data from a 46-h experiment, circles for data from a 168-h experiment), and for molten mid-ocean-ridge basalt (diamonds for data from a 264-h experiment). The isotopic fractionations are given as $\delta^{26}\mathrm{Mg}(‰) = \left(\frac{(^{26}\mathrm{Mg}/^{24}\mathrm{Mg})_{\mathrm{sample}}}{(^{26}\mathrm{Mg}/^{24}\mathrm{Mg})_{\mathrm{starting\ material}}} - 1\right) \times 1{,}000$. The figure shows that in the case of andesite, Ω_{Mg} is not constant, but rather decreases by a factor of more than two between $T > 1{,}500$ °C and $T < 1{,}500$°C. The data for basalt are too few to support a conclusion to the effect that Ω_{Mg} is independent of composition.

Huang *et al.*[3] claim to have cast their experimental observations in a theoretical framework. This 'theoretical framework' is simply a parameterization of an assumed, not theoretically derived, functional form for the mass dependence of a quantity they call ΔS_{T}, which, like Ω, is proportional to the slope of the thermal isotope fractionations versus temperature. The assumed expression for ΔS_{T} is $\Delta S_{\mathrm{T}} = c\left(\frac{Z^2}{a}\right)^{\beta}(X^{\alpha} - Y^{\alpha})$, where X and Y are the mass of two isotopes of an element of valence Z and ionic radius a. The three quantities c, α and β were adjusted such that the calculated ΔS_{T} reproduces three values of ΔS_{T} taken from their experiments. Leaving aside the issues of what specific value of ΔS_{T} to choose, given its variation with temperature, which is proportional to the slope of δ^{26}Mg versus temperature shown in Fig. 1, or that all the elements considered have the same Z, using a parameterization with three free parameters to fit three data does not constrain and validate the functional form of the parameterization. One can, however, test the proposed parameterization by calculating ΔS_{T} for all the major elements of basalt (magnesium, calcium, iron, silicon and oxygen) and comparing these to the experimental data reported by Richter *et al.*[2]. As might be expected, the parameterization works reasonably well in reproducing the measured ΔS_{T} for magnesium, calcium and iron; however, when applied to oxygen and silicon, the calculated values and the measured values differ by more than a factor two for oxygen and more than a factor of ten for silicon.

I caution against assuming (on the basis of the data presented by Huang *et al.*[3]) that thermal isotope fractionations in silicate liquids are independent of temperature and composition or that they have found "a simple and robust framework for characterizing isotope fractionations by thermal diffusion in natural and synthetic systems".

Frank M. Richter[1]
[1]Department of the Geophysical Sciences, The University of Chicago,
Chicago, Illinois 60637, USA.
e-mail: richter@geosci.uchicago.edu

Received 20 April 2010; accepted 14 January 2011.

1. Richter, F. M., Watson, E. B., Mendybaev, R. A., Teng, F.-Z. & Janney, P. E. Magnesium isotope fractionation in silicate melts by chemical and thermal diffusion. *Geochim. Cosmochim. Acta* **72,** 206–220 (2008).
2. Richter, F. M. *et al.* Isotope fractionation of the major elements of molten basalt by chemical and thermal diffusion. *Geochim. Cosmochim. Acta* **73,** 4250–4263 (2009).
3. Huang, F. *et al.* Isotope fractionation in silicate melts by thermal diffusion. *Nature* **464,** 396–400 (2010).

Competing financial interests: declared none.

doi:10.1038/nature09954

©2011 Macmillan Publishers Limited. All rights reserved

Huang *et al.* reply

REPLYING TO F. M. Richter *Nature* **472,** doi:10.1038/nature09954 (2011)

In our Letter[1], we showed that the phenomenon of isotope fractionation in silicate melts by thermal diffusion, first reported in 1998 (ref. 2), can be characterised by a parameter ΔS_T that is independent of temperature and composition. (Here ΔS_T is the difference in the Soret coefficient, S_T, between isotopes of a diffusing element.) Richter[3] questioned this finding by plotting Mg isotope ratio versus temperature data (figure 1 in ref. 3) for a subset of experiments from figure 3f of ref. 1.

Using the same coordinate system as Richter (figure 1 of ref. 3), here in Fig. 1a we plot the Mg data for all of the experiments (figure 3f of ref. 1), and in Fig. 1b we plot the Fe data for all of the experiments (figure 3d of ref. 1). Focussing on individual experiments, we find that isotope fractionation, as measured by the local slope of the data, may increase (as emphasised by Richter[3]) or decrease, but commonly follow a near-linear distribution. Further, when the same data are plotted as isotope ratios versus relative temperature (ΔT), as was done in ref. 1,

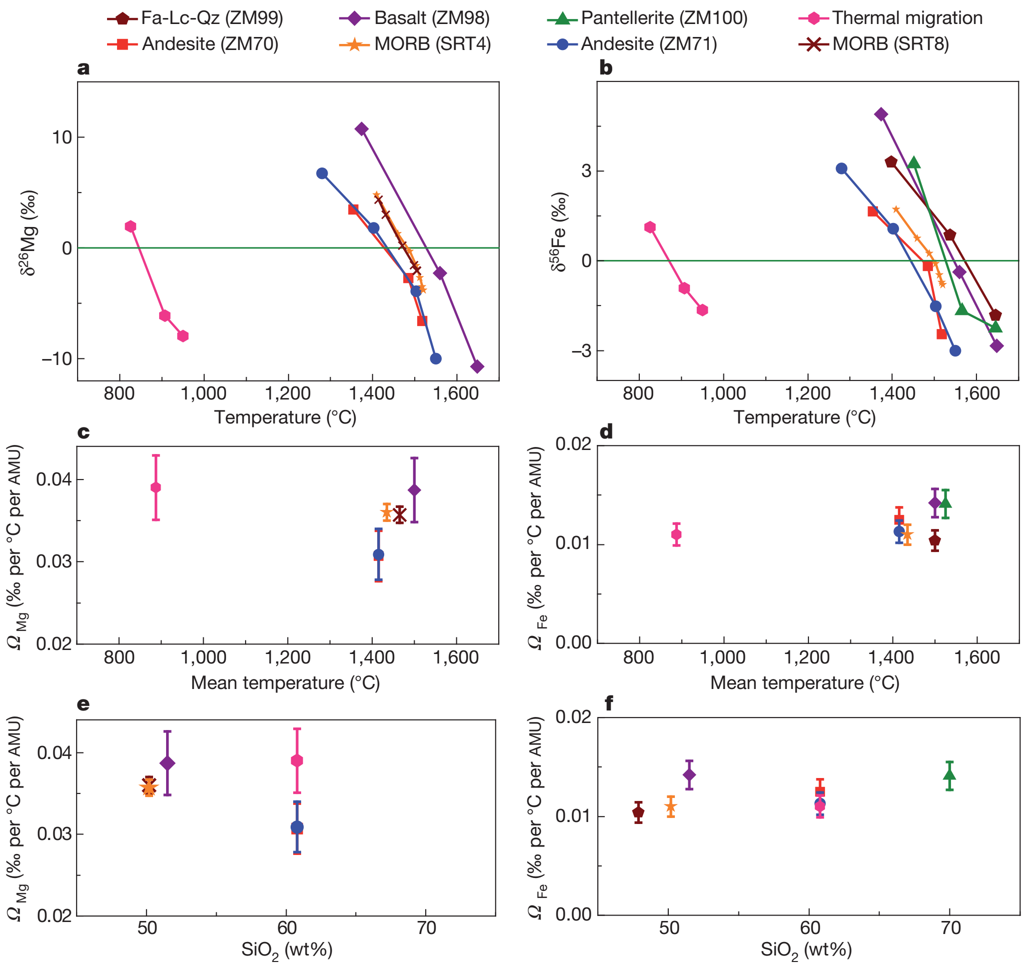


Figure 1 | Variation of Mg and Fe isotope data with composition and temperature. a, b, Mg (**a**) and Fe (**b**) isotope ratio versus temperature. The same data are presented as isotope ratio versus relative temperature (ΔT) in figure 3d, f of ref. 1. The isotope ratios are expressed as $\delta^{26}Mg = 1{,}000 \times [(^{26}Mg/^{24}Mg)_{sample}/(^{26}Mg/^{24}Mg)_{DSM\text{-}3} - 1]$ (‰) and $\delta^{56}Fe = 1{,}000 \times [(^{56}Fe/^{54}Fe)_{sample}/(^{56}Fe/^{54}Fe)_{IRMM\text{-}14} - 1]$ (‰). **c, d,** Mg (**c**) and Fe (**d**) thermal-diffusion isotopic sensitivity (Ω, expressed in units of ‰ $°C^{-1}$ AMU^{-1}) versus mean temperature; **e, f,** Ω_{Mg} (**e**) and Ω_{Fe} (**f**) versus SiO_2 content of the experimental starting materials. The slope of isotope ratio versus temperature data for Mg and Fe (in **a** and **b**) is proportional to Ω_{Mg} and Ω_{Fe} (in **c**–**f**), respectively. Here (and in figure 1 of ref. 1) we compute the Ω (and the corresponding error bar, ±1 s.d., calculated on the basis of errors in isotope analyses and temperature measurements) for each experiment using the data from the cold and hot ends. Note that despite wide variations in experimental conditions, the isotope ratio versus temperature profiles are approximately linear, and the Ω values are approximately independent of temperature and composition. The same conclusion, albeit based on a limited range of experimental conditions, also applies for Ca (ref. 1). The data sources for the experiments are: ref. 1 for Fe_2SiO_4-$KAlSi_2O_6$-SiO_2 (Fa-Lc-Qz), pantellerite, andesite and basalt; ref. 4 for thermal migration; and refs 9 and 10 for MORB. The thermal migration experiment was conducted using USGS andesite standard AGV-1 (plus 4% H_2O) as starting material at 5 kbar and in a temperature gradient from 950 to 350 °C (ref. 4). Thermal diffusion experiments of ref. 1 were carried out at 10 kbar using starting materials with SiO_2 content ranging from 47.9 to 70 wt% (refs 11, 12). (As an aside, we note a labelling error in figure 3b of ref. 1: the red curve corresponds to ZM70 and the blue curve corresponds to ZM71.)

©2011 Macmillan Publishers Limited. All rights reserved

all experiments to first order collapse onto a linear distribution (with the coefficient of determination, R^2, being 0.94 for Mg data and 0.93 for Fe data; see figure 3d and f in ref. 1); this is despite wide variations in starting composition (48–70 wt% SiO_2), mean temperature (850–1,525 °C), and in one case, the presence of co-existing mineral phases[4]. These findings cannot be reconciled if there are significant temperature and compositional dependences on isotope fractionation arising from thermal diffusion.

Moreover, it is ill-advised to attach much significance to incremental changes in slope along a thermal diffusion profile composed of relatively few, widely spaced analyses, and with non-trivial temperature uncertainties for the intermediate positions as compared to the cold and hot ends of the temperature gradient. If, instead, one considers the thermal-diffusion isotopic sensitivity (Ω), which reflects the overall fractionation between the hot and cold ends, the differences for Mg and Fe (Fig. 1c–f), as well as for Ca, are small and unsystematic with temperature or composition—observations that reinforce our original claim.

Richter[3] also maintains that our theoretical treatment is little more than data fitting. This view seems to miss the significance of the linear relationships shown in figure 3d–f of ref. 1, which demonstrate no significant variation in ΔS_T with temperature or composition. These observations enabled the additive decomposition of S_T (equation (5) in ref. 1), which provides a strong constraint for a general theory of thermal diffusion in silicate melts. It further permits links to thermal diffusion in other systems[5–8]. The fit that we presented in the Supplementary Information of ref. 1 was simply an example of using equation (5)[1] to seek empirical formulae for ΔS_T using the available data for network modifiers.

Last, we agree with Richter[3] that our treatment is not suitable for silicon and oxygen, but we find his criticism misplaced, as we explicitly stated in our letter[1] that "We limit our attention to the steady state of thermal diffusion and, furthermore, to the isotopic fractionation of iron, calcium and magnesium (which break up the polymerization of the silicate melt and are thus termed network modifiers), and do not consider network formers (for example silicon and oxygen, which form tetrahedron networks in silicate melts)."

F. Huang[1,4], P. Chakraborty[1], C. C. Lundstrom[1], C. Holmden[2], J. J. G. Glessner[3], S. W. Kieffer[1] & C. E. Lesher[3]
[1]Department of Geology, University of Illinois, Urbana, Illinois 61801, USA.
[2]Department of Geological Sciences, University of Saskatchewan, Saskatoon, Saskatchewan S7N 5E2, Canada.
[3]Department of Geology, University of California, Davis, California 95616, USA.
[4]CAS Key Laboratory of Crust-Mantle Materials and Environments, School of Earth and Space Sciences, University of Science and Technology of China, Hefei 230026, China.
e-mail: fang.huang@erdw.ethz.ch and fhuang@ustc.edu.cn

1. Huang, F. *et al.* Isotope fractionation in silicate melts by thermal diffusion. *Nature* **464,** 396–400 (2010).
2. Kyser, T. K., Lesher, C. E. & Walker, D. The effects of liquid immiscibility and thermal diffusion on oxygen isotopes in silicate liquids. *Contrib. Mineral. Petrol.* **133,** 373–381 (1998).
3. Richter, F. M. Isotope fractionation in silicate melts by thermal diffusion. *Nature* **472,** doi:10.1038/nature09954 (this issue).
4. Huang, F. *et al.* Chemical and isotopic fractionation of wet andesite in a temperature gradient: experiments and models suggesting a new mechanism of magma differentiation. *Geochim. Cosmochim. Acta* **73,** 729–749 (2009).
5. Astumian, R. D. Coupled transport at the nanoscale: the unreasonable effectiveness of equilibrium theory. *Proc. Natl Acad. Sci. USA* **104,** 3–4 (2007).
6. Debuschewitz, C. & Kohler, W. Molecular origin of thermal diffusion in benzene + cyclohexane mixtures. *Phys. Rev. Lett.* **87,** 055901 (2001).
7. Putnam, S. A., Cahill, D. G. & Wong, G. C. L. Temperature dependence of thermodiffusion in aqueous suspensions of charged nanoparticles. *Langmuir* **23,** 9221–9228 (2007).
8. Reith, D. & Muller-Plathe, F. On the nature of thermal diffusion in binary Lennard-Jones liquids. *J. Chem. Phys.* **112,** 2436–2453 (2000).
9. Richter, F. M. *et al.* Magnesium isotope fractionation in silicate melts by chemical and thermal diffusion. *Geochim. Cosmochim. Acta* **72,** 206–220 (2008).
10. Richter, F. M. *et al.* Isotopic fractionation of the major elements of molten basalt by chemical and thermal diffusion. *Geochim. Cosmochim. Acta* **73,** 4250–4263 (2009).
11. Lesher, C. E. & Walker, D. Solution properties of silicate liquids from thermal diffusion experiments. *Geochim. Cosmochim. Acta* **50,** 1397–1411 (1986).
12. Lesher, C. E. & Walker, D. in *Diffusion, Atomic Ordering, and Mass Transport* (ed. Ganguly, J.) 396–451 (Adv. Phys. Geochem. 8, Springer, 1991).

Competing financial interests: declared none.

doi:10.1038/nature09955

©2011 Macmillan Publishers Limited. All rights reserved